الجمهورية الجزائرية الديمقراطية الشعبية
République algérienne démocratique et populaire
وزارة التعليم العالي والبحث العلمي
Ministère de l'enseignement supérieur et de la recherche scientifique
جامعة عين تموشنت بلحاج بوشعيب
Université –Ain Temouchent- Belhadj Bouchaib
Faculté des Sciences et de Technologie
Département de Mathématiques et Informatique

Projet de Fin d'Etudes
Pour l'obtention du diplôme de Master en Mathématiques

Domaine : Mathématiques et Informatique
Filière : Mathématiques
Spécialité : Probabilités et Statistique Appliquées

Thème

Prévision des Séries Chronologiques à l'aide de la Régression Linéaire

Présentée par :
Mlle OURRAG Nourhane Keltoum

Devant le jury composé de :

Dr. SAKHI Hanane	MCB	UAT.B.B (Ain Temouchent)	Présidente
Dr. MECENE Rahmouna	MCB	UAT.B.B (Ain Temouchent)	Examinatrice
Dr. BENNAFLA Djamila	MCB	UAT.B.B (Ain Temouchent)	Encadrante

Année Universitaire 2024/2025

Résumé

Ce mémoire explore l'utilisation de la régression linéaire pour la prévision des séries chronologiques. Nous analysons des séries temporelles réelles et mettons en œuvre des techniques de régression linéaire pour prédire les valeurs futures. Nous accordons une attention particulière à l'analyse des données, au diagnostic des modèles et à l'évaluation de la performance prédictive. Nous proposons également des exemples pratiques et des simulations sous R, afin de permettre au lecteur de comprendre les concepts et d'appliquer les méthodes présentées.

Summary

This thesis explores the use of linear regression for time series forecasting. We analyze real-world time series data and implement linear regression techniques to predict future values. We pay special attention to data analysis, model diagnostics, and the evaluation of predictive performance. We also include practical examples and simulations using R, enabling the reader to understand the concepts and apply the presented methods.

ملخص

يتناول هذا البحث استخدام الانحدار الخطي في التنبؤ بالسلاسل الزمنية. نقوم بتحليل بيانات لسلاسل زمنية حقيقية ، ونطبق تقنيات الانحدار الخطي لتقدير القيم المستقبلية. نولي اهتمامًا خاصًا بتحليل البيانات، وتشخيص النماذج ، وتقييم الأداء التنبؤي. كما نقدم أمثلة تطبيقية و محاكاة باستخدام برنامج R، لتمكين القارئ من فهم المفاهيم و تطبيق الأساليب المعروضة.

Dédicace

À ma chère famille,

En témoignage de ma profonde gratitude pour votre amour,
votre patience et votre soutien indéfectible tout au long de mon parcours.

Ce mémoire vous est dédié, en reconnaissance de votre présence perpétuelle,
de vos encouragements silencieux et de votre foi en moi,
même dans les moments les plus difficiles.

Sans vous, rien de tout cela n'aurait été possible.

Keltoum Nourhane OURRAG

Remerciements

Avant toute chose, je tiens à exprimer ma plus profonde gratitude et ma reconnaissance éternelle au Tout-Puissant Allah, qui m'a guidée et m'a donné la force nécessaire pour surmonter tous les obstacles et les difficultés rencontrés tout au long de ce parcours de recherche en master.

Je remercie sincèrement ma chère encadrante, ***Dr. D. BENNAFLA****, pour son soutien indéfectible et son aide précieuse. Sans son accompagnement académique, ce mémoire n'aurait jamais pu voir le jour, ni même être mené à terme. Je ne saurais exprimer avec justesse toute l'estime et la reconnaissance que j'ai pour sa présence bienveillante. Ma gratitude envers elle est immense, et je lui en suis profondément reconnaissante.*

J'adresse également mes remerciements sincères aux membres du jury, ***Dr. H. SAKHI*** *et* ***Dr. R. MECENE****, pour m'avoir fait l'honneur de leur présence et pour avoir accepté d'évaluer ce travail.*

Enfin, je souhaite aussi exprimer ma reconnaissance à ma chère sœur ***Ghizlene*** *et à la personne qui m'est la plus précieuse sur le plan personnel et moral dans mon parcours,* ***Hbiba****, pour leurs soutiens sincères et leurs encouragements chaleureux durant les moments difficiles de la réalisation de ce mémoire.*

À tous ceux qui ont contribué, de près ou de loin, à la réalisation de ce travail, je dis simplement : ***merci****.*

Table des matières

Liste des tableaux

Table des figures

Notations et abréviations

Ω	Univers des événements aléatoires
$\mathcal{F}$	Tribu définie sur Ω
P	Mesure de probabilité définie sur $\mathcal{F}$
$\mathbb{R}$	Ensemble des nombres réels
$\mathbb{Z}$	Ensemble des nombres entiers relatifs
$\mathbb{N}$	Ensemble des nombres entiers naturels
M.M	Méthode des moyennes mobiles
MAE	Erreur absolue moyenne
MSE	Erreur quadratique moyenne
RLS	Régression linéaire simple
RLM	Régression linéaire multiple
BB	Bruit blanc
MCO	Moindres Carrés Ordinaires
IC	Intervalle de Confiance

Introduction

Dans un monde où l'information évolue en continu et où les données temporelles sont omniprésentes, la capacité à anticiper l'avenir à partir de l'observation du passé s'impose comme un enjeu fondamental dans de nombreux domaines : économie, agriculture, météorologie, énergie, santé publique, entre autres. Cette démarche est connue sous le nom de prévision des séries chronologiques.

Une **série chronologique** est une suite d'observations réalisées à intervalles de temps réguliers, reflétant l'évolution d'un phénomène dans le temps. Cette évolution peut être influencée par plusieurs composantes, telles qu'une tendance de fond, des effets saisonniers, des cycles économiques ou encore des variations aléatoires.

Comprendre ces composantes permet non seulement d'analyser le comportement passé du phénomène, mais aussi d'en anticiper l'évolution. C'est dans cette perspective que s'inscrit la **prévision**, dont l'objectif principal est d'estimer les valeurs futures à partir des données historiques, tout en réduisant au maximum l'incertitude associée.

Pour atteindre cet objectif de prévision, il est essentiel de s'appuyer sur des modèles capables de capturer les structures sous-jacentes des séries temporelles. C'est dans cette optique que la **modélisation des séries chronologiques** a été progressivement développée au fil du temps. Elle remonte au début du XX^{e} siècle avec les premiers travaux de Yule (1927), Slutsky (1937), puis surtout avec les contributions fondamentales de Box et Jenkins (1970), qui ont posé les bases des modèles **ARIMA**. Depuis, de nombreuses méthodes ont vu le jour, intégrant des approches plus sophistiquées telles que les modèles **SARIMA**, les modèles à composantes multiples, ou encore les **réseaux de neurones** appliqués à la prévision.

Au-delà des modèles ARIMA et de leurs extensions, plusieurs approches empiriques ont été développées, parmi lesquelles les *moyennes mobiles*, le *lissage exponentiel* et la *décomposition des séries*. Ces méthodes permettent, chacune selon ses spécificités, d'extraire les composantes essentielles des séries temporelles afin d'en faciliter l'analyse et la prévision. Toutefois, la régression linéaire demeure l'un des outils les plus accessibles, interprétables et efficaces, notamment lorsqu'il s'agit d'identifier et de modéliser une tendance linéaire ou une structure saisonnière régulière dans les données.

Si les méthodes classiques de prévision des séries temporelles sont puissantes, elles restent parfois lourdes à mettre en œuvre et nécessitent des conditions strictes, comme la stationnarité ou l'absence d'autocorrélation résiduelle. Dans ce cadre, la régression linéaire apparaît comme un outil plus souple et plus intuitif. Ce mémoire se propose donc de répondre à la problématique suivante :

Dans quelle mesure la régression linéaire, sous ses différentes formes (simple, multiple, temporelle), constitue-t-elle une méthode pertinente et efficace pour la prévision des séries chronologiques, notamment lorsqu'elles présentent des tendances marquées ou des schémas saisonniers ?

Le présent travail a pour objectifs de :

— Comprendre les concepts fondamentaux des séries chronologiques et leurs propriétés.

— Présenter le cadre théorique de la régression linéaire appliqué aux données temporelles.

— Développer et valider des modèles de régression pour la prévision de séries réelles.

— Comparer la performance de différents types de régressions : simple, multiple, et temporelle.

— Mettre en lumière les avantages et les limites de la méthode.

Pour répondre à cette problématique et atteindre les objectifs fixés, ce mémoire s'articule autour de trois chapitres principaux, construits de manière progressive afin de passer des fondements théoriques aux applications pratiques :

Chapitre 1 : Concepts de la prévision pour les séries chronologiques.

Ce chapitre introduit les notions fondamentales de séries chronologiques, telles que la stationnarité, l'autocorrélation, la structure temporelle et les composantes des séries. Il présente également un aperçu des méthodes classiques de prévision.

Chapitre 2 : Cadre théorique de la prévision par la régression linéaire.

Ce chapitre est dédié à l'étude des modèles de régression linéaire dans le contexte temporel. Il aborde les hypothèses, l'ajustement du modèle, les critères d'évaluation et les méthodes de validation, avec une attention particulière portée à l'interprétation des résultats.

Chapitre 3 : Application sur données réelles.

Ce dernier chapitre applique les modèles développés sur deux jeux de données temporelles. Il présente les étapes de visualisation, de modélisation, d'évaluation et de prévision. Une attention est portée à la comparaison entre les prévisions issues de la régression simple et multiple.

Le mémoire se conclut par une synthèse des principaux résultats, mettant en évidence les apports de la **régression linéaire** à la prévision des séries chronologiques. Il propose également quelques pistes de réflexion pour de futurs approfondissements, et s'accompagne d'une **bibliographie** regroupant l'ensemble des sources mobilisées.

Chapitre 1

Concepts de la prévision pour les séries chronologiques

Ce chapitre introduit les notions fondamentales liées aux séries chronologiques et à la prévision. Il présente les principales propriétés statistiques des séries temporelles, ainsi que les différentes méthodes de prévision et les fondements théoriques sur lesquels elles reposent.

Les éléments abordés ici constituent la base du travail développé dans les chapitres suivants et s'appuient sur des références théoriques reconnues dans le domaine, notamment : [3] [4] [6] [8] [10] [12] [14] et [15].

1.1 Mécanismes temporels dans les phénomènes aléatoires

Toute analyse rigoureuse des séries chronologiques repose sur une compréhension préalable des notions de base. Nous commencerons donc par définir les concepts de *processus stochastique* et de *série chronologique*.

1.1.1 Processus stochastiques

Définition 1.1.1 *Un* ***processus stochastique*** Y *est une collection de variables aléatoires* $\{Y_t; t \in T\}$*, où* T *est un ensemble d'indices pour lequel toutes les* Y_t *sont définies sur un même espace probabilisé* $(\Omega, \mathcal{F}, P)$*.*

Remarque 1.1.1 *Lorsque* T *représente le temps, on appelle le processus une* ***série temporelle****, tel que :*

- $\forall\ t \in T$ *fixé,* Y_t *est une variable aléatoire.*
- $\forall\ \omega \in \Omega$ *fixé,* $t \mapsto Y_t(\omega)$ *est une fonction du temps à valeurs réelles, appelée trajectoire du processus* Y*.*

1.1.2 Séries chronologiques

Définition 1.1.2 *On appelle une* ***série chronologique*** *(ou* ***série temporelle****) toute suite finie réelle* $(Y_t)_{t\in T}$ *d'observations numériques d'un phénomène aléatoire ordonnées au cours du temps, généralement relevées à des intervalles équidistants.*

$$Y_t = D_t + N_t,$$

où

- Y_t *est la valeur de la* $t^{ième}$ *observation de la série temporelle.*
- D_t *est la composante déterministe.*
- N_t *est la composante aléatoire .*

Dans la suite, L'axe temporel T peut être :

- ***Discret*** *: si $T = \mathbb{Z}$.*
- ***Continu*** *: si $T = \mathbb{R}$.*

Remarque 1.1.2 *Une serie chronologique n'est donc qu'une realisation d'un processus stochastique.*

Exemple 1.1.1 *La figure suivante représente le nombre annuel de taches solaires observées à la surface du Soleil entre les années 1700 et 1980, illustrant ainsi l'évolution de cette activité au fil du temps.*

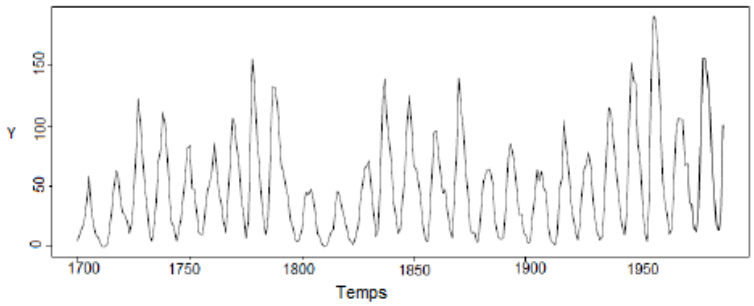


FIGURE 1.1 – Exemple de série chronologique.

Afin de mieux appréhender le comportement d'une série chronologique, il convient d'en examiner les propriétés fondamentales, qui permettent de caractériser sa dynamique dans le temps.

1.2 Propriétés de base

1.2.1 La stationnarité

La stationnarité constitue une propriété clé des séries chronologiques, car elle garantit que les caractéristiques statistiques du processus ne varient pas au fil du temps.

Définition 1.2.1 (Stationnarité stricte) *Un processus $\{Y_t; t \in T\}$ est dit* ***strictement stationnaire*** *si, pour toute famille finie d'instants $t_1, t_2, \ldots, t_k \in T$ avec $k \in \mathbb{N}^*$, et pour tout décalage $h \in T$, la distribution jointe de $\{Y_{t_1}, \ldots, Y_{t_k}\}$ est identique à celle de $\{Y_{t_1+h}, \ldots, Y_{t_k+h}\}$.*

$$\mathcal{L}(Y_{t_1}, \ldots, Y_{t_k}) \overset{d}{=} \mathcal{L}(Y_{t_1+h}, \ldots, Y_{t_k+h}).$$

Définition 1.2.2 (Stationnarité en covariance) *Une série temporelle $Y = (Y_t)_{t\in T}$ est dit* ***stationnaire en covariance*** *si :*

(i) $\mathbb{E}(Y_t) = \mu, \quad \forall t \in T.$

(ii) $\mathbb{E}(Y_t^2) = \sigma^2 < \infty, \quad \forall t \in T.$

(iii) $\mathrm{Cov}(Y_t, Y_{t+h}) = \gamma(h), \quad$ *(dépend uniquement de h et non de t).*

Remarque 1.2.1 *La stationnarité en covariance est également appelée stationnarité faible, stationnarité au sens large ou stationnarité d'ordre 2.*

Propriété 1.2.1 *Si un processus aléatoire est fortement (ou strictement) stationnaire, alors il est stationnaire au second ordre. Par contre, la réciproque n'est pas toujours vraie.*

La vérification de la stationnarité ouvre la voie à l'analyse des relations temporelles dans la série, notamment à travers les fonctions d'autocovariance, d'autocorrélation et d'autocorrélation partielle, qui permettent de décrire plus finement les liens entre les valeurs passées et présentes de la série.

1.2.2 Fonction d'autocovariance

Définition 1.2.3 *Si $\{Y_t; t \in T\}$ une série temporelle stationnaire, alors pour tout $t \in T$, la fonction d'autocovariance $\gamma(\cdot)$ définie par :*

$$\forall h \in T, \quad \gamma(h) = \mathbb{E}[(Y_t - \mu)(Y_{t+h} - \mu)],$$

satisfait les propriétés suivante :

(i) $\gamma(0) = \sigma^2$.

(ii) $|\gamma(h)| \leq \gamma(0), \quad \forall h \in T$.

(iii) $\gamma(h) = \gamma(-h), \quad \forall h \in T$.

(iv) La fonction d'autocovariance $\gamma(h)$ est associée à une matrice symétrique semi-définie positive, comme dans l'expression suivante :

$$\gamma(h) = \begin{pmatrix} \gamma(0) & \gamma(1) & \dots & \gamma(h-1) \\ \gamma(1) & \gamma(0) & \dots & \gamma(h-2) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(h-1) & \gamma(h-2) & \dots & \gamma(0) \end{pmatrix}, \quad \forall h \in T.$$

1.2.3 Fonction d'autocorrélation

Définition 1.2.4 *Si $Y = (Y_t)_{t \in T}$ un processus stationnaire, la fonction d'autocorrélation $\rho(\cdot)$ définie par :*

$$\forall h \in T, \quad \rho(h) = \mathrm{Corr}(Y_t, Y_{t+h}) = \frac{\gamma(h)}{\gamma(0)},$$

satisfait les propriétés analogues suivantes :

(i) $\rho(0) = 1$.

(ii) $|\rho(h)| \leq 1, \quad \forall h \in T$.

(iii) $\rho(h) = \rho(-h), \quad \forall h \in T$.

(iv) La fonction d'autocorrélation $\rho(h)$ donne lieu à la matrice suivante, caractérisée par sa symétrie et sa semi-définie positivité :

$$\rho(h) = \begin{pmatrix} 1 & \rho(1) & \cdots & \rho(h-1) \\ \rho(1) & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho(h-1) & \cdots & \cdots & 1 \end{pmatrix}, \quad \forall h \in T.$$

Théorème 1.2.1 *Si $\gamma(0) > 0$ et $\lim_{j \to \infty} \gamma_j = 0$, alors $\gamma(h)$ et $\rho(h)$ sont définies positives pour chaque entier $h > 0$.*

Remarque 1.2.2 *Les fonctions $\gamma(h)$ et $\rho(h)$ permettent de mesurer le degré de dépendance entre les valeurs d'une série temporelle à différents instants. Pour cette raison, elles ont un rôle important lorsqu'il s'agit de prévoir les valeurs futures de la série en fonction des valeurs passées et présentes.*

1.2.4 Fonction d'autocorrélation partielle

L'autocorrélation partielle mesure la corrélation entre deux observations Y_t et Y_{t+h}, sans prendre en considération l'influence des variables intermédiaires, peut être calculée à l'aide du rapport suivant :

$$\phi(h) = \frac{|\rho(h^*)|}{|\rho(h)|}; \quad \forall h \in \mathbb{N}^*,$$

où

$$\rho(h^*) = \begin{pmatrix} 1 & \rho(1) & \cdots & \rho(1) \\ \rho(1) & 1 & \cdots & \rho(2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(h-1) & \rho(h-2) & \cdots & \rho(h) \end{pmatrix},$$

et

$$\rho(h) = \begin{pmatrix} 1 & \rho(1) & \cdots & \rho(h-1) \\ \rho(1) & 1 & \cdots & \rho(h-2) \\ \vdots & \vdots & \ddots & \vdots \\ \rho(h-1) & \rho(h-2) & \cdots & 1 \end{pmatrix}.$$

Ici, $|\rho(h)|$ et $|\rho(h^*)|$ désignent respectivement les déterminants des matrices carrées $\rho(h)$ et $\rho(h^*)$. Les trois premières autocorrélations partielles sont déterminées par les relations :

$$\phi(1) = \rho(1),$$

$$\phi(2) = \frac{\rho(2) - \rho(1)^2}{1 - \rho(1)^2},$$

$$\phi(3) = \frac{\rho(1)^3 - \rho(1)\rho(2)(2 - \rho(2)) + \rho(3)(1 - \rho(1)^2)}{1 - \rho(2)^2 - 2\rho(1)^2(1 - \rho(2))}.$$

Remarque 1.2.3 *La fonction d'autocorrélation partielle permet d'isoler l'effet direct d'une observation passée sur la valeur actuelle, ce qui facilite une meilleure compréhension des relations temporelles importantes pour améliorer la qualité des prévisions.*

Pour illustrer ces propriétés fondamentales des séries temporelles, un exemple représentatif est présenté afin de mieux comprendre leur signification et leur manifestation dans les données.

Exemple 1.2.1 *Le bruit blanc est l'exemple le plus simple, il fait partie de la classe des processus stationnaires. Spécifiquement, $(\varepsilon_t)_t$ est un bruit blanc si :*

$$\mathbb{E}(\varepsilon_t) = 0, \quad \mathbb{E}(\varepsilon_t^2) = \sigma_\varepsilon^2 < \infty, \quad \textit{et} \quad \rho(h) = \mathrm{Corr}(\varepsilon_t, \varepsilon_{t+h}) = 0, \quad \forall h \geq 1.$$

À travers l'exemple, nous illustrons graphiquement la série temporelle, en mettant en évidence ses caractéristiques à l'aide des graphes ACF et PACF.

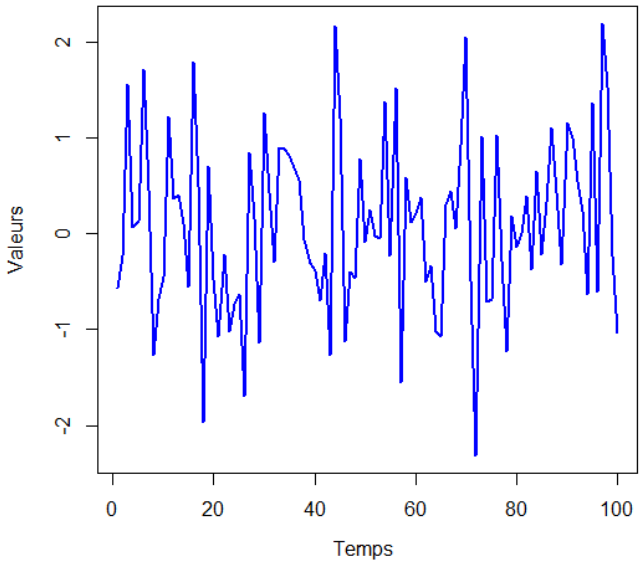


FIGURE 1.2 – Trajectoire d'un bruit blanc.

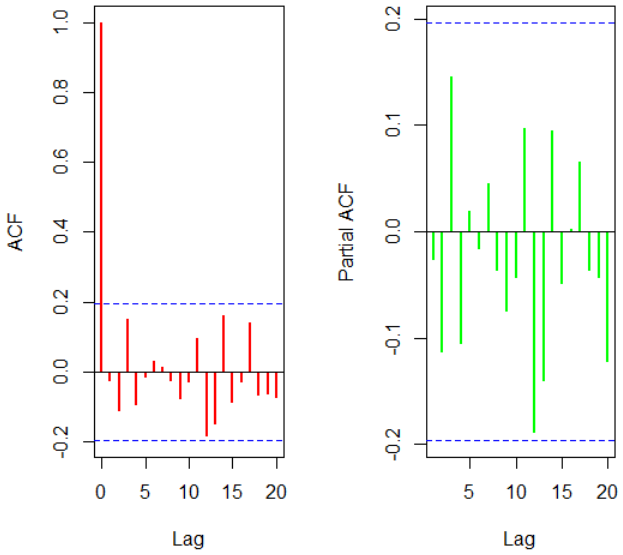


FIGURE 1.3 – Fonctions d'autocorrélation et partielle d'un bruit blanc.

À l'observation de la trajectoire du bruit blanc, la série oscille autour d'un niveau constant avec une amplitude relativement stable, ce qui suggère que ses propriétés statistiques restent homogènes au cours du temps.

L'analyse des fonctions d'autocorrélation confirme cette observation : la première barre atteint la valeur maximale, reflétant une autocorrélation parfaite d'une observation avec elle-même, tandis que les valeurs suivantes restent très faibles et majoritairement comprises dans les intervalles de confiance (représentés par les bandes bleues). Ceci traduit l'absence de corrélations significatives entre les observations et confirme le caractère aléatoire de la série.

Une fois les propriétés des séries chronologiques analysées, il est essentiel de caractériser leur structure, de distinguer leurs différentes composantes, et d'orienter le choix du modèle de prévision le plus adapté.

1.3 Structure des séries chronologiques

1.3.1 Composantes fondamentales

On considère qu'une série chronologique est la résultante de différentes composantes fondamentales qui gouvernent son comportement au fil du temps.

- **La tendance τ_t (trend) :**

Cette composante représente l'évolution « moyenne » à long terme de la variable étudiée . Il faut noter que la notion de « long terme » est peu précise et varie selon le problème posé.

Cela se traduit concrètement par l'exemple suivant, qui montre une trajectoire de série caractérisée par une tendance.

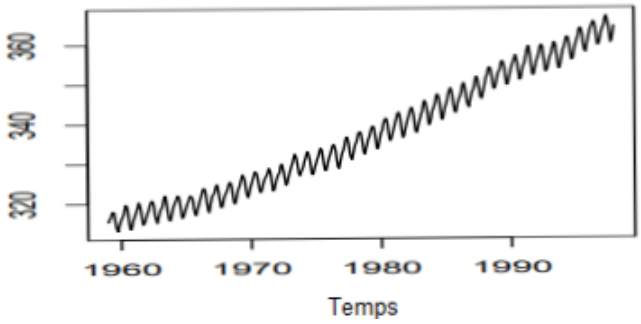


FIGURE 1.4 – Exemple de trajectoire d'une série affichant une tendance.

- **La saisonnalité S_t (Seasonality) :**

Cette caractéristique d'une série chronologique correspond à un phénomène saisonnière qui se répète à intervalles de temps réguliers (périodiques) de p période .

Cela se traduit concrètement par l'exemple suivant, qui illustre une série présentant une saisonnalité clairement identifiable.

$$\forall t \in T, \quad S_{t+p} = S_t.$$

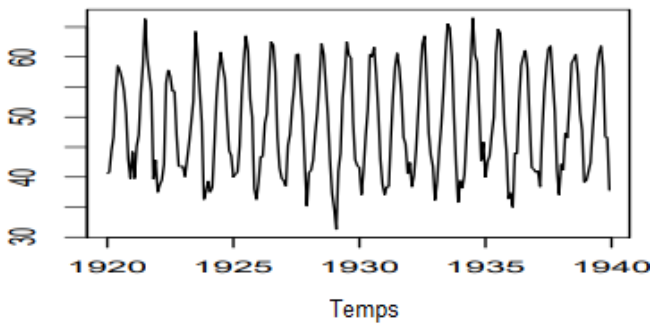


FIGURE 1.5 – Exemple de trajectoire d'une série saisonnière.

- **Les résidus ε_t (residuals) :**

Les mouvements irréguliers (ou résiduelles) sont des fluctuations accidentelles et aléatoires. Ils sont généralement impossibles à prévoir et peuvent résulter d'événements imprévus. On parle aussi d'aléas.

Cela se traduit concrètement par l'exemple suivant, qui illustre une série comportant des variations irrégulières imprévisibles.

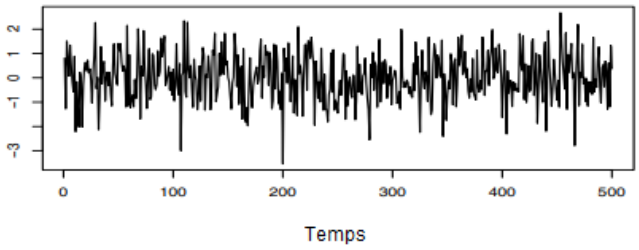


FIGURE 1.6 – Exemple de trajectoire d'une série irrégulier.

- **Les Cycles C_t (Cycls) :**

Les mouvements cycliques sont des fluctuations qui se présentent sur une période plus ou moins longue autour de la courbe de tendance, souvent difficile à définir. Ces variations ne sont reconnues comme cycliques que si elles se reproduisent après une période d'au moins une année. Les étapes du cycle sont rarement complètement uniformes car les conditions historiques (les causes) qui les provoquent changent beaucoup d'une étape d'un cycle à une autre.

Cela se traduit concrètement par l'exemple suivant, qui illustre une série présentant des fluctations cycliques sur une période étendu.

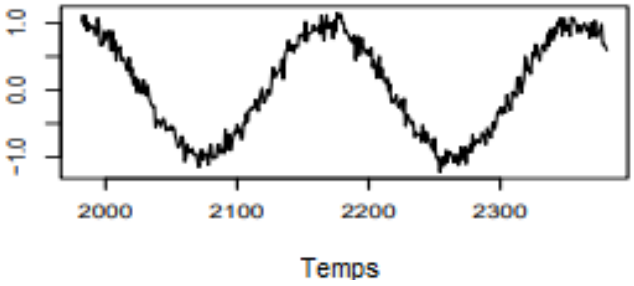


FIGURE 1.7 – Exemple de trajectoire d'une série affichant un cycle.

La connaissance des composantes principales d'une série chronologique nous permet de la décomposer efficacement. Cette décomposition facilite grandement l'analyse en isolant clairement chacune de ces composantes, ce qui aide à mieux comprendre la dynamique globale de la série.

À titre d'exemple, considérons la série représentant le nombre mensuel de passagers dans un aéroport sur une période donnée, qui illustre bien ces différentes composantes et leur impact sur les données, comme présenté dans la figure 1.8.

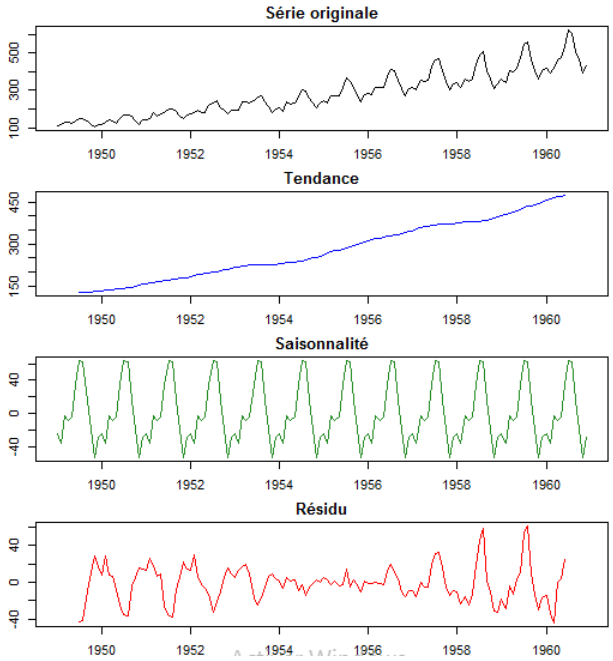


FIGURE 1.8 – Décomposition de la série chronologique.

La compréhension des composantes d'une série chronologique ouvre naturellement la voie à sa modélisation. Dans la section suivante, nous nous concentrons sur les modèles additif et multiplicatif, qui permettent de représenter la structure de la série à partir de ses composantes identifiées, en mettant l'accent sur le choix entre ces deux formulations selon la nature des données.

1.3.2 Modélisation

Un modèle est une représentation simplifiée de la réalité, permettant d'expliquer le fonctionnement du phénomène étudié et de mieux comprendre ses composantes ainsi que leurs interactions. Une fois un modèle est obtenu, il peut etre utilisé pour la prédiction des valeurs futurs.

- **Le modèle additif :**

C'est le **"modèle classique de décomposition"** dans le traitement des modèles d'ajustement des séries chronologiques. la variable d'intérêt Y_t se décompose sous la forme suivante :

$$Y_t = \tau_t + S_t + \varepsilon_t,$$

où $(\varepsilon_t)_t$ la composante aléatoire indépendante et identiquement distribuée (i.i.d).

Le modèle additif est le plus approprié si l'ampleur des fluctuations saisonnières, ou la variation autour de la tendance-cycle, ne varie pas avec le niveau de la série chronologique.

- **Le modèle multiplicatif :**

Dans ce cas, on parle de modèle multiplicatif lorsque la variable Y_t est exprimée sous la forme :

$$Y_t = \tau_t \times S_t \times \varepsilon_t.$$

Le modèle multiplicatif convient plus lorsque la variation de la tendance saisonnière, ou la variation autour de la tendance-cycle, semble proportionnelle au niveau de la série chronologique.

Après avoir présenté les modèles additif et multiplicatif, il convient désormais de s'interroger sur le modèle le plus approprié à la structure de la série étudiée. Cette étape de choix est déterminante pour assurer la pertinence de l'analyse et la qualité des prévisions.

1.3.3 Choix du modèle

À cet égard, un exemple de données d'une série chronologique présenté dans la Figure 1.9 met en évidence que l'amplitude des variations reste constante autour de la tendance.

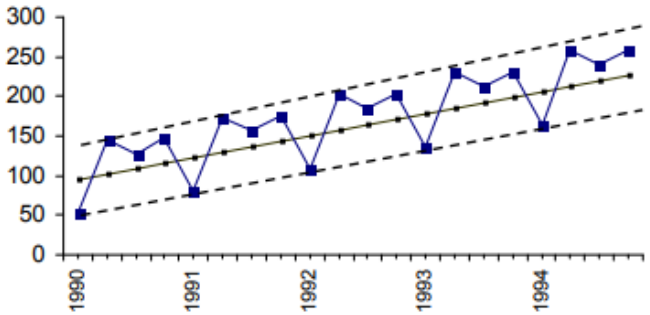


FIGURE 1.9 – Exemple d'une série additif.

On remarque que les deux droites tracées sont à peu près parallèles entre elles et gardent la même amplitude au long du temps, ce qui nous permet de dire que le modèle est additif.

Encore, un autre exemple de données d'une série chronologiques présenté dans la Figure 1.10 montre que l'amplitude des variations saisonnières varie.

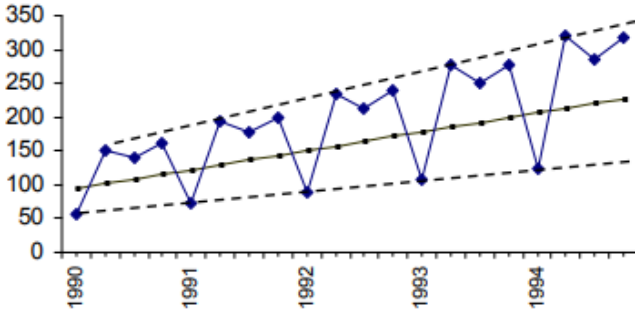


FIGURE 1.10 – Exemple d'une série multiplicatif.

Dans ce cas, l'absence de parallélisme entre les deux droites et la croissance de leur amplitude au fil du temps permettent de dire que le modèle est multiplicatif.

Après avoir construit des modèles capables de représenter les composantes principales d'une série chronologique, il devient pertinent d'exploiter ces modèles pour anticiper l'évolution future des données. C'est dans cette optique que s'inscrit la prévision, qui constitue une finalité essentielle de l'analyse des séries chronologiques.

1.4 Prévisions des séries chronologiques

La prévision consiste à estimer les valeurs futures d'une série à partir des observations passées, en tenant compte des régularités détectées dans les données. Elle s'appuie sur les résultats de la modélisation pour prolonger la dynamique observée, selon un horizon temporel défini (court, moyen ou long terme), en fonction des besoins de l'étude.

Dans ce travail, seules les séries chronologiques à intervalles réguliers seront considérées (par exemple : mensuelles, trimestrielles, annuelles). Les séries à intervalles irréguliers, bien qu'elles existent dans certaines applications, ne seront pas traitées ici car elles nécessitent des approches spécifiques.

La mise en place d'une stratégie de prévision repose sur le choix de méthodes adaptées aux caractéristiques de la série. Nous présentons ici les principales approches, en mettant en lumière leurs objectifs et les types de données auxquelles elles conviennent le mieux.

1.4.1 Méthodes de prévision

Les méthodes de prévision des séries chronologiques peuvent être classées en plusieurs grandes familles, selon leur manière de traiter les données et de prendre en compte les composantes de la série. Parmi les approches les plus courantes, on distingue :

— Les moyennes mobiles, qui visent à lisser les fluctuations pour dégager une tendance locale ;

— Le lissage exponentiel, qui accorde plus de poids aux observations récentes pour capturer la dynamique actuelle ;

— La décomposition saisonnière, qui sépare les différentes composantes pour mieux comprendre leur contribution.

Chaque méthode a ses spécificités et son champ d'application privilégié, selon la structure de la série et les objectifs de prévision (réactivité, stabilité, robustesse...).

- **Méthode des Moyennes Mobiles :**

La moyenne mobile est une méthode simple permettant d'extraire la composante tendance d'une série temporelle. Elle est également connue comme une méthode de lissage utilisée pour éliminer les fluctuations à court terme.

Soit Y une série chronologique portant sur n années, décomposée en p saisons, et soit $l \in \{2, \dots, np\}$. On appelle *moyenne mobile* (ou glissante) d'ordre l au temps $t \in \left[1 + \left\lfloor \frac{l}{2} \right\rfloor, \dots, np - \left\lfloor \frac{l}{2} \right\rfloor\right]$ où $\lfloor \cdot \rfloor$ désigne la partie entière, la quantité :

$$M_t^{(l)} = \begin{cases} \dfrac{1}{l}\left(\dfrac{Y_{t-\frac{l}{2}}}{2} + Y_{t-\frac{l}{2}+1} + \cdots + Y_t + \cdots + Y_{t+\frac{l}{2}-1} + \dfrac{Y_{t+\frac{l}{2}}}{2}\right), & \text{si } l \text{ est pair,} \\ \dfrac{1}{l}\left(Y_{t-\frac{l-1}{2}} + \cdots + Y_t + \cdots + Y_{t+\frac{l-1}{2}}\right), & \text{si } l \text{ est impair.} \end{cases}$$

Lorsque la moyenne mobile donne une prévision stable, cela indique que la série est probablement stable ; lorsqu'elle augmente, cela reflète une tendance haussière récente, mais avec un certain retard de réaction ; et lorsqu'il existe un écart important entre la prévision et la valeur réelle, cela signifie que la méthode est trop simple pour capturer des dynamiques complexes comme une tendance marquée ou une saisonnalité.

- **Méthode Lissage exponentiel :**

Le lissage exponentiel ne repose sur aucune théorie statistique formelle, mais relève plutôt de la pratique. Il est utilisé avec les séries chronologiques dont la moyenne change lentement dans le temps.

1. Lissage exponentiel simple :

Le lissage exponentiel simple s'applique à des séries chronologiques sans tendance et sans saisonnalité. Le modèle prend la forme suivante :

$$\hat{y}_t = ay_{t-1} + (1-a)\hat{y}_{t-1},$$

où $\hat{y}_t$ la prévision pour t, $\hat{y}_{t-1}$ est la valeur observée précédent, et $a \in [0,1]$ le paramètre de lissage.

Remarque 1.4.1 *Pour choisir le paramètre a, on peut utiliser trois méthodes :*

- ***Essais et erreurs** : tester plusieurs valeurs entre 0 et 1 et comparer les résultats (visuellement ou via des critères d'erreur).*
- ***Minimisation d'un critère d'erreur** : choisir a qui minimise l'erreur quadratique moyenne (MSE) ou l'erreur absolue moyenne (MAE) sur un échantillon d'apprentissage.*
- ***Règles empiriques** : souvent, a est choisi entre 0,1 et 0,3 pour des séries avec peu de bruit, et plus élevé (jusqu'à 0,5 voire 0,8) si la série est volatile.*

La valeur prévisionnelle obtenue à partir du lissage exponentiel simple représente une estimation pondérée qui tient davantage compte des observations récentes, tout en conservant une mémoire des valeurs passées. Cette prévision reflète l'évolution récente de la série, avec une réactivité modulée par le paramètre de lissage a. Si a est élevé, la prévision est fortement influencée par la dernière observation, ce qui la rend plus sensible aux variations récentes. En revanche, si a est faible, la prévision s'appuie davantage sur les tendances passées, produisant une estimation plus lissée mais moins réactive. Ainsi, la prévision obtenue donne une valeur attendue au prochain instant, en supposant que la dynamique récente se poursuit, ce qui en fait un outil pertinent pour les séries stables et non saisonnières à court terme.

2. Lissage exponentiel double (Holt) :

Le lissage exponentiel de Holt s'applique aux séries chronologiques sans composante saisonnière et à tendance localement linéaire. Le modèle est défini par le système suivant :

$$\begin{cases} \hat{\nu}_t = ay_t + (1-a)(\hat{\nu}_{t-1} + \hat{\tau}_{t-1}), \\ \hat{\tau}_t = b(\hat{\nu}_t - \hat{\nu}_{t-1}) + (1-b)\hat{\tau}_{t-1}, \\ \hat{y}_{t+h} = \hat{\nu}_t + h\hat{\tau}_t. \end{cases}$$

où $\hat{\nu}_t$ représente le niveau estimé à l'instant t, $\hat{\tau}_t$ la tendance estimée à l'instant t, $\hat{y}_{t+h}$ la prévision à l'horizon h, et $(a,b) \in [0,1]$ les paramètres de lissage.

Lorsque la prévision obtenue par lissage exponentiel double est proche de la dernière valeur observée, cela indique que la série est stable, avec peu de variations récentes et une évolution quasi constante. Si la prévision est supérieure à la dernière valeur, le modèle met en évidence une tendance haussière, suggérant une augmentation attendue de la variable. En revanche, si elle est inférieure, cela traduit une tendance baissière que le modèle anticipe comme se prolongeant. Une prévision très différente des valeurs récentes peut révéler soit une tendance forte nouvellement

apparue, soit un ajustement excessif dû à des paramètres de lissage trop élevés, ou encore une inadéquation du modèle face à une série comportant des saisonnalités ou des ruptures. Enfin, si la tendance estimée est proche de zéro, le modèle interprète la série comme étant quasiment stationnaire, malgré la prise en compte théorique d'une composante de tendance.

3. Lissage exponentiele triple (Holt-Winter) :

Le lissage exponentiel de Holt-Winters s'applique aux séries chronologiques a à la fois une tendance et une saisonnalité.

Formules additive : si les effets saisonniers sont constants.

$$\begin{cases} \hat{\nu}_t = a(y_t - S_{t-p}) + (1-a)(\hat{\nu}_{t-1} + \hat{\tau}_{t-1}), \\ \hat{\tau}_t = b(\hat{\nu}_t - \hat{\nu}_{t-1}) + (1-b)\hat{\tau}_{t-1}, \\ S_t = c(y_t - \hat{\nu}_t) + (1-c)S_{t-p}, \\ \hat{y}_{t+h} = \hat{\nu}_t + h\hat{\tau}_t + S_{t+h-p}. \end{cases}$$

Formules multiplicative : si les effets saisonniers varient selon le niveau de la série.

$$\begin{cases} \hat{\nu}_t = a\left(\frac{y_t}{S_{t-p}}\right) + (1-a)(\hat{\nu}_{t-1} + \hat{\tau}_{t-1}), \\ \hat{\tau}_t = b(\hat{\nu}_t - \hat{\nu}_{t-1}) + (1-b)\hat{\tau}_{t-1}, \\ S_t = c\left(\frac{y_t}{\hat{\nu}_t}\right) + (1-c)S_{t-p}, \\ \hat{y}_{t+h} = (\hat{\nu}_t + h\hat{\tau}_t) \times S_{t+h-p}. \end{cases}$$

où S_t est la saisonnalité, p la périodicité et $(a, b, c) \in [0, 1]$.

Avec le lissage exponentiel triple, la prévision intègre le niveau, la tendance et la saisonnalité. Si elle est proche des valeurs récentes, la série est stable avec une saisonnalité régulière. Une prévision plus élevée indique une hausse attendue, due à une tendance positive ou à une phase saisonnière haute, tandis qu'une prévision plus basse signale une baisse prévue. Un écart important avec les valeurs récentes peut refléter une saison forte/faible, une tendance récente marquée ou un mauvais ajustement. Si la tendance et la saisonnalité sont faibles, la série est considérée comme quasiment stationnaire.

- **Méthode de décomposition :**

Il n'y a pas de théorie statistique à proprement parler derrière les méthodes de décomposition, elles relèvent plutôt de l'intuition et de la pratique. Elles sont employées avec les séries qui montrent une tendance et une composante saisonnière.

1. Nous commençons par estimer la tendance τ_t, on utilisera donc un filtre selon la période de la saisonnalité. Cela s'effectue à l'aide d'un filtre de M.M finie avec :

$$\hat{\tau}_t = \begin{cases} \frac{1}{l}\left(\frac{1}{2}Y_{t-q} + Y_{t-q+1} + \cdots + Y_t + \cdots + Y_{t+q-1} + \frac{1}{2}Y_{t+q}\right), & \text{si } l = 2q, \\ \frac{1}{l}\left(Y_{t-q} + \cdots + Y_t + \cdots + Y_{t+q}\right), & \text{si } l = 2q+1. \end{cases}$$

où $q + 1 < t \leq n - q, \quad \forall q \in \mathbb{N}^*$.

2. Ainsi, pour estimer la composante saisonnière, il convient d'abord d'étudier deux cas :

*** Cas 1 : Modèle additif**

— On calcule les données sans tendance $Y_t - \tau_t$,

— On calcule la moyenne des données sans tendance du mois j sur les n années, ceci pour chacun des p périodes. D'où :

$$S_j = \frac{1}{n}\sum_{i=1}^{n}(Y_{ij} - \tau_{ij}), \quad \forall j = 1, 2, \ldots, p.$$

— On calcule la moyenne des coefficients saisonniers S_j :

$$\bar{S} = \frac{1}{p}\sum_{j=1}^{p} S_j.$$

— Si $\bar{S} \neq 0$, on corrige les coefficients saisonniers CS_j :

$$CS_j = S_j - \bar{S}.$$

*** Cas 2 : Modèle multiplicatif**

— On calcule les données sans tendance $\frac{Y_t}{\tau_t}$,

— On calcule la moyenne des données sans tendance du mois j sur les n années, ceci pour chacun des p périodes :

$$S_j = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_{ij}}{\tau_{ij}}, \quad \forall j = 1, 2, \ldots, p.$$

— On calcule la moyenne des coefficients saisonniers S_j :

$$\bar{S} = \frac{1}{p}\sum_{j=1}^{p} S_j.$$

— Si $\bar{S} \neq 1$, on corrige les coefficients saisonniers CS_j :

$$CS_j = \frac{S_j}{\bar{S}}.$$

3. Ensuite, pour la désaisonnalitée de la série :

— **Modèle additif :** $\overset{*}{Y} = Y_{ij} - CS_{ij}$,

— **Modèle multiplicatif :** $\overset{*}{Y} = \frac{Y_{ij}}{CS_{ij}}$.

4. Enfin, on ajuste la tendance en prenant la série désaisonnalisée, avec la méthode des moindres carrés en l'appliquant aux points $(t, \overset{*}{Y}_t)$, où $\overset{*}{Y}_t$ s'exprime selon la nature de la tendance :

— Si la tendance est **linéaire** :

$$\overset{*}{Y}_t = \beta_0 + \beta_1 t,$$

— Si la tendance est **quadratique** :

$$\overset{*}{Y}_t = \beta_0 + \beta_1 t + \beta_2 t^2.$$

Lorsqu'on utilise la méthode de décomposition, plusieurs situations peuvent se présenter. Si la tendance est stable et la saisonnalité régulière, la prévision est considérée comme fiable. Une tendance croissante indique une augmentation progressive de la variable, tandis qu'une tendance décroissante reflète une baisse attendue. Une forte saisonnalité amplifie ou atténue ces tendances selon la période considérée. En revanche, si les résidus sont élevés ou irréguliers, cela signifie que le modèle n'explique pas bien une partie des variations, ce qui peut être dû à des facteurs imprévus ou à un mauvais ajustement.

1.4.2 Limites et avantages des méthodes de prévision

1. **Moyennes mobiles :**
   Les moyennes mobiles, simples à utiliser, lissent les fluctuations courtes, mais elles ne prennent pas en compte la tendance ni la saisonnalité, ce qui limite leur efficacité.

2. **Lissage exponentiel :**
   Le lissage exponentiel simple accorde plus de poids aux données récentes, mais reste inefficace en présence de tendance ou de saisonnalité. Le lissage exponentiel double (Holt) améliore cette méthode en prenant en compte la tendance, toutefois il ne gère pas la saisonnalité. Enfin, le lissage exponentiel triple (Holt-Winters) modélise à la fois tendance et saisonnalité, mais son paramétrage est plus complexe et il est sensible aux changements brusques des données.

3. **Méthode de décomposition :**
   La méthode de décomposition permet une analyse claire des composantes de la série. Cependant, elle devient moins fiable lorsque la saisonnalité ou les résidus présentent une instabilité importante, ce qui peut affecter la qualité de la modélisation et des prévisions.

Les méthodes de lissage et de décomposition permettent d'identifier les composantes fondamentales d'une série chronologique, telles que la tendance et la saisonnalité, en s'appuyant uniquement sur les valeurs passées de la variable étudiée. Toutefois, ces approches restent limitées lorsqu'il s'agit de comprendre ou de prédire une série influencée par des facteurs extérieurs.

Dans ce contexte, la régression linéaire constitue une alternative méthodologique puissante, en permettant de modéliser explicitement la relation entre la variable à prévoir et une ou plusieurs variables explicatives.

La régression linéaire simple est particulièrement adaptée aux séries présentant une tendance linéaire, en prenant le temps comme prédicteur. En présence de saisonnalité, la régression linéaire multiple permet d'intégrer des variables indicatrices représentant les fluctuations périodiques. En plus de fournir des prévisions, cette approche permet une interprétation des relations entre les variables, ce qui en fait un outil à la fois explicatif et prédictif.

Le chapitre suivant sera ainsi consacré à la présentation détaillée de la régression linéaire appliquée à la prévision des séries chronologiques, en tant que prolongement naturel des méthodes classiques.

Chapitre 2

Cadre théorique de prévision des séries chronologiques par la régression

Dans ce chapitre, nous présentons les modèles de régression linéaire appliqués à la prévision des séries chronologiques. Le principe fondamental repose sur l'hypothèse que la série étudiée entretient une relation linéaire avec une ou plusieurs autres variables temporelles. Ce cadre théorique permet d'expliquer et de prédire l'évolution de la série à partir de facteurs explicatifs.

Les modèles de prévision peuvent être de nature **déterministe** ou **stochastique**. Les premiers décrivent l'évolution de la série par une relation mathématique explicite, tandis que les seconds introduisent une composante aléatoire dans la dynamique des données. Ce mémoire se limite au cadre déterministe, en particulier à l'utilisation de la **régression linéaire**.

Pour enrichir le contenu de ce chapitre, nous nous sommes appuyés sur plusieurs références spécialisées, notamment : [2] [5] [7] [8] [9] [10] [12] [13] [14] et [17].

2.1 Modèles de régression linéaire

La régression linéare est une technique statistique parmi les méthodes les plus importantes en mathématiques appliquées, employée pour décrire ou expliquer la relation entre les variables dans le but d'estimer les paramètres d'un modèle ou, plus précisément, pour prédire la valeur d'une variable **endogène** (à expliquer) Y en fonction des variables **exogènes** (explicatives) X.

L'étude des relations entre les variables fait largement appel à des modèles construits sur la base de théories existantes. La spécification du modèle est un ensemble d'hypothèses sur la forme générale de la relation que l'on étudie, et sur les caractéristiques des éléments aléatoires qui y figurent. La statistique a alors pour rôle :

— Estimer les paramètres du modèle,

— Indiquer la confiance que l'on peut mettre dans ces estimés et de tester les hypothèses que l'on peut faire.

C'est ainsi que se développe l'étude statistique des modèles de régression linéaire .

Le modèle de régression linéaire est le modèle de prévision le plus courant pour identifier la relation entre une variable dépendante et des variables indépendantes. Hormis les types de données univariées ou multivariées, le concept est linéaire. La régression linéaire peut être soit simple, soit multiple.

2.1.1 Régression linéaire simple RLS

Définition 2.1.1 *Le modèle simple permet une relation linéaire entre la variable de prévision y et une seule variable prédictive x comme dans la relation :*

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \forall i = 1, \ldots, n.$$

où

— β_0 *: Coefficient désignant l'ordonnée à l'origine.*

— β_1 *: Pente de la droite.*

— ε_i *: le terme d'erreur aléatoire.*

Quand $x = 0$, β_0 *représente la valeur prédite de* y.

Un exemple simulé de données provenant d'un tel modèle est présenté dans la Figure 2.1.

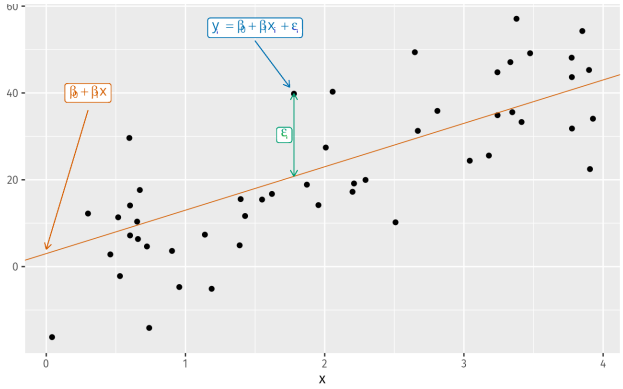


FIGURE 2.1 – Exemple d'un modèle de régression linéaire simple.

2.1.2 Régression linéaire multiple RLM

Définition 2.1.2 *Lorsqu'il existe deux variables prédictives ou plus, le modèle est linéaire multiple présentée dans la relation suivante de la forme générale :*

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i, \quad \forall i = 1, \ldots, n, \quad \forall k \in \mathbb{N}^*.$$

où y_i *est la variable dépendante, et* $x_{i1} \ldots, x_{ik}$ *sont les variables indépendantes d'ordre* k. *les coefficients* $\beta_1, \ldots, \beta_k$ *mesurent l'effet indépendant de chaque prédicteur, après avoir pris en compte les effets de tous les autres prédicteurs du modèle.*

Ainsi, la Figure 2.2 correspond également à un autre exemple simulé de données multiples issues d'un tel modèle.

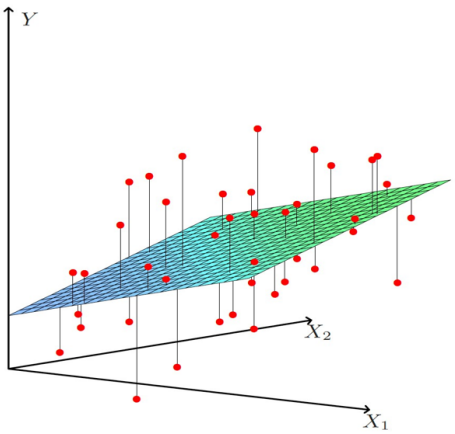


FIGURE 2.2 – Exemple d'un modèle de régression linéaire multiple.

Après avoir présenté la régression linéaire dans son cadre simple et multiple, nous allons maintenant aborder son extension au contexte des séries chronologiques. Précisons que la régression linéaire simple est le cas particulier où il n'y a qu'une seule variable explicative. Dans certains cas, Nous aborderons ici le cas général de la régression linéaire multiple.

2.1.3 Régression linéaire temporelle RLT

Définition 2.1.3 *Le modèle temporel représente une relation linéaire entre une série temporelle dépendante, disons y_t, pour $t = 1, \ldots, L$, et un ensemble de séries indépendantes, disons $x_{t1}, x_{t2}, \ldots, x_{tk}$. cette relation s'écrit sous la forme suivante :*

$$y_t = \beta_0 + \beta_1 x_{t1} + \beta_2 x_{t2} + \cdots + \beta_k x_{tk} + \varepsilon_t, \quad \forall t = 1, \ldots, L, \quad \forall k \in \mathbb{N}^*.$$

où $\beta_0, \beta_1, \ldots, \beta_k$ sont des coefficients de régression fixes mais inconnus, et $(\varepsilon_t)_t$ est un processus d'erreur ou de bruit aléatoire (i.i.d) $\sim \mathcal{N}(0, \sigma_\epsilon^2)$.

2.1.4 Hypothèses du modèle

Lorsque nous utilisons un modèle de régression linéaire, nous faisons implicitement certaines hypothèses sur les erreurs :

(H1) : $\mathbb{E}(\varepsilon_t) = 0 \quad \forall t \in T$, (Centrage).

(H2) : $\mathrm{Var}(\varepsilon_t) = \sigma^2$, (Homoscédasticité).

(H3) : $\mathrm{Cov}(\varepsilon_t, \varepsilon_{t'}) = 0, \quad \forall t \neq t' \in T$, (Indépendance).

(H4) : $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$, (Normalité).

(H5) : $\mathrm{Cov}(x_{tk}, \varepsilon_t) = 0, \quad \forall k \in \mathbb{N}^*$, (Indépendance entre x_{tk} et ε_t).

Le respect de ces hypothèses est indispensable pour appliquer la régression linéaire conventionnelle. Une fois ces conditions réunies, il devient possible d'estimer les coefficients du modèle à l'aide de la méthode des moindres carrés, largement utilisée dans ce cadre.

Remarque 2.1.1

1. Dans la régression appliquée aux séries temporelles, il est rare que le bruit soit blanc. Il sera donc nécessaire, à terme, d'assouplir cette hypothèse.

2. Les résidus présentent certaines propriétés utiles, comme nous l'avons déjà vu, notamment :

$$\sum_{t=1}^{L} \varepsilon_t = 0, \quad et \quad \sum_{t=1}^{L} x_{tk}\, \varepsilon_t = 0, \quad \forall k \in \mathbb{N}^*.$$

Ces égalités ne sont pas nécessairement vérifiées lorsque l'ordonnée à l'origine est omise dans le modèle.

Après avoir présenté les modèles de régression linéaire, il est nécessaire d'examiner leurs différentes composantes afin de mieux comprendre leur structure interne et leurs mécanismes de fonctionnement.

2.2 Analyse des composantes du modèle linéaire

La série Y_t est la somme de deux composantes déterministes : une tendance τ_t, d'une saisonnalité S_t, et d'une composante aléatoire ε_t :

$$Y_t = \tau_t + S_t + \varepsilon_t,$$

On suppose que τ_t et S_t sont des combinaisons linéaires de fonctions connues dans le temps, τ_t^i et S_t^j, i.e.

$$\begin{cases} \tau_t = \tau_t^1\beta_1 + \tau_t^2\beta_2 + \cdots + \tau_t^m\beta_m, \\ S_t = S_t^1\lambda_1 + S_t^2\lambda_2 + \cdots + S_t^p\lambda_p. \end{cases} \quad \forall\beta_i \in \mathbb{R}, \quad \forall\lambda_j \in \mathbb{R}.$$

Le but est d'estimer les paramètres $\beta_1, \ldots, \beta_m$ et $\lambda_1, \ldots, \lambda_p$ à partir des L observations.

L'expression complète devient :

$$Y_t = \sum_{i=1}^{m} \tau_t^i\beta_i + \sum_{j=1}^{p} S_t^j\lambda_j + \varepsilon_t, \quad \forall t = 1, \ldots, L.$$

où m est le nombre de composantes de tendance et p le nombre de composantes saisonnières.

Remarque 2.2.1 *Dans ce modèle, chaque période saisonnière est représentée par une variable indicatrice spécifique. Ainsi, comme un paramètre est estimé pour chaque période, le nombre de composantes saisonnières est égal à la périodicité.*

- **Composante saisonnière du modèles :**

La forme de S_t dépend du type de données et de la forme de la saisonnalité. On considérera ici des fonctions indicatrices S_t^i, définies comme suit :

$$S_t^i = \begin{cases} 1 & \text{si } t = \text{mois } i, \\ 0 & \text{sinon.} \end{cases}$$

Exemple 2.2.1 *Pour des données trimestrielles, on a :*

$$S_t = S_t^1\lambda_1 + S_t^2\lambda_2 + S_t^3\lambda_3 + S_t^4\lambda_4,$$

où S_t^j est la fonction indicatrice du trimestre j, c'est-à-dire :

$$S_t^j = \begin{cases} 1 & \textit{si } t \textit{ est dans le trimestre } j, \\ 0 & \textit{sinon.} \end{cases}$$

- **Composante tendancielle :**

Il est courant que les données de séries chronologiques présentent des tendances. Plusieurs types de composantes tendancielles existent :

(i) Linéaire : $\tau_t = \beta_0 + \beta_1 t$.

(ii) Exponentielle : $\tau_t = \alpha\beta^t$, ou $\tau_t = \alpha(1+\beta)^t$, ou encore $\tau_t = \alpha\exp(\beta t)$.

(iii) Quadratique : $\tau_t = \beta_0 + \beta_1 t + \beta_2 t^2$.

(iv) De Gompertz : $\tau_t = \exp(\alpha\beta^t + \lambda)$.

(v) Logistique : $\tau_t = [\alpha\beta^t - \lambda]^{-1}$.

Remarque 2.2.2 *Le cas (i) se traite par régression simple (cf partie suivante), Le cas (ii) peut être ramené au cas (i) par transformation logarithmique, Le cas (iii) se traite par régression multiple. Il est également possible d'utiliser des modèles avec ruptures :*

$$\tau_t = \begin{cases} \alpha_0 + \alpha_1 t & \text{pour } t \leq t_0, \\ \beta_0 + \beta_1 t & \text{pour } t > t_0. \end{cases}$$

Cette forme de tendance est l'une des plus difficiles à modéliser, car il n'existe pas vraiment de méthode universelle.

Exemple 2.2.2 *Considérons comme variable le logarithme de l'indice de la Bourse de New York, représenté ci-dessous, sur laquelle nous avons tenté trois ajustements différents :* ***linéaire***, ***quadratique***, ***exponentiel***.

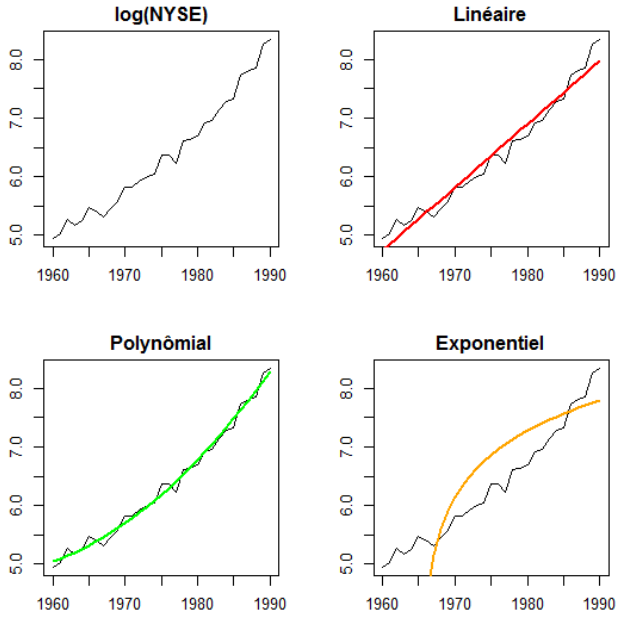


FIGURE 2.3 – L'indice de la Bourse de New York.

Remarque 2.2.3 *La vraie difficulté est que la modélisation doit pouvoir avoir un "sens" : une tendance linéaire indique une croissance linéaire, alors qu'une tendance exponentielle indique une augmentation constante. En revanche, une tendance quadratique peut être plus difficile à justifier puisque la plupart des modèles structurels sont généralement additifs (linéaires) ou multiplicatifs (linéaires après transformation logarithmique). Les tendances linéaires avec rupture sont également très utilisées, puisqu'elles sont souvent plus adaptées qu'une tendance linéaire "simple", et surtout, la rupture a une interprétation structurelle.*

2.2.1 Analyse de tendance linéaire

— Nous nous plaçons dans le cadre d'un modèle composé uniquement d'une tendance et de fluctuations irrégulières et donnons une méthode permettant d'estimer cette tendance.

— Une tendance linéaire peut être modélisée simplement en utilisant $x_{t1} = t$ en tant que prédicteur.

— Lorsqu'une liaison linéaire forte entre une varible y et une date t semble raisonnable au vu du nuage de points :

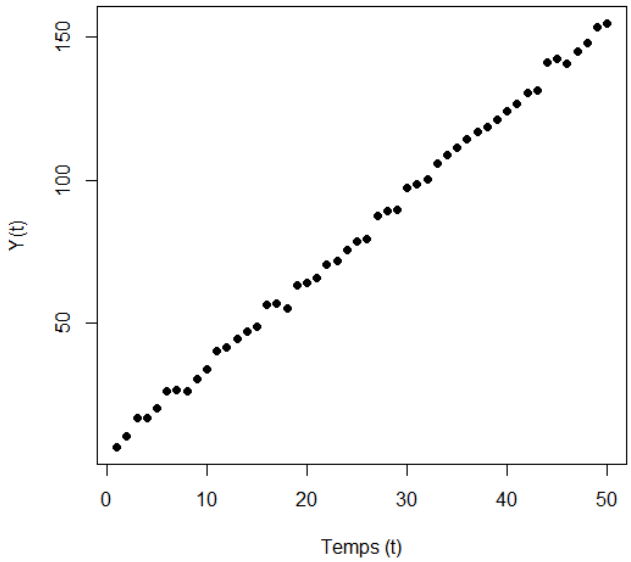


FIGURE 2.4 – Exemple de nuage de points illustrant une tendance linéaire.

On a alors une relation du type :

$$y_t = \tau_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad \forall t = 1, \ldots, L.$$

• Le problème ici est donc d'éstimer ces coefficients grace à valeur observer sur l'échantillon t. On cherche ainsi, à déterminer la droite qui s'ajuste le mieux aux donnée c'est à dire la droite la plus proche des points du nuage. Pour cela, il faut donc mesurer l'éloignement des points du nuage par rapport à une droite (D) d'équation $y = \beta_1 t + \beta_0$ puis minimiser un critère d'erreur donné.

• Pour estimer l'équation de la tendance nous régresson la série sur le temps avec la méthode des moindes carrées (MCO), on doit minimiser l'erreur quadratique moyenne relative aux résidus du modèle, ce qui s'écrit :

$$g(\beta_0, \beta_1) = \sum (y - \beta_1 t - \beta_0)^2.$$

Pour cela, on cherche β_0 et β_1 tels que :

$$\frac{\partial g(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \text{et} \quad \frac{\partial g(\beta_0, \beta_1)}{\partial \beta_1} = 0,$$

et on trouve dans ce cas le couple solution $(\hat{\beta}_0, \hat{\beta}_1)$ donné par :

$$\hat{\beta}_1 = \frac{\sum (t - \bar{t})(y - \overline{y})}{\sum (t - \bar{t})^2} = \frac{\mathrm{Cov}(t, y_t)}{\mathrm{Var}(t)} = \frac{\overline{ty} - \bar{t}\,\overline{y}}{\overline{t^2} - \bar{t}^2}, \quad \text{et} \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1\,\bar{t},$$

avec

$$\bar{t} = \frac{1}{L}\sum t, \quad \overline{y} = \frac{1}{L}\sum y, \quad \text{et} \quad \overline{t^2} = \frac{1}{L}\sum t, \quad \overline{ty} = \frac{1}{L}\sum ty.$$

où $L \in \mathbb{N}^*$ est le nombre d'observations.

• La droite d'équation $y = \hat{\beta}_1 t + \hat{\beta}_0$ est appelée droite de régression de y en t notée : $\Delta_{y/t}$.

- La droite de régression moyenne quant à elle s'écrit :

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{t}.$$

On utilise l'équation de la régression pour prévoir les valeurs de la variable dépendante pour d'autres valeurs des variables indépendantes.

2.2.2 Analyse de tendance non linéaire

La méthode la plus simple pour modéliser une relation non linéaire consiste à transformer la variable explicative avant d'estimer un modèle de régression. Bien que cela fournisse une forme fonctionnelle non linéaire, le modèle reste linéaire dans ses paramètres.

La transformation la plus couramment utilisée est la transformation logarithmique :

$$\log(y_t) = \beta_0 + \beta_1 \log(x_t) + \varepsilon_t.$$

Exemple 2.2.3

— *Si* $\tau_t = \beta_1 t^2 + \beta_0$*, en posant* $y_t = t^2$*, on se ramène à* $\tau_t = \beta_1 y_t + \beta_0$*, et on peut faire un ajustement linéaire entre* y_t *et* τ_t*.*

— *Si* $\tau_t = \beta_0 \exp(\beta_1 t)$*, en posant* $y_t = \ln(\tau_t)$*, on se ramène à* $y_t = \beta_1 t + \ln(\beta_0)$*, et on peut faire un ajustement linéaire entre* y_t *et* t*.*

L'ensemble des éléments abordés dans cet exemple est représenté de manière explicite dans le graphique 2.3, offrant ainsi une lecture visuelle cohérente de l'analyse.

2.2.3 Analyse de saisonalité

1. Modèle trimestriel de Buys-Ballot (1847) :

Le modèle de régression linéaire, dans le cas où la tendance est supposée linéaire, et les données sont trimestrielles s'écrit comme suit :

$$Y_t = \underbrace{\beta_0 + \beta_1 t}_{\tau_t} + \underbrace{\lambda_1 S_t^1 + \lambda_2 S_t^2 + \lambda_3 S_t^3 + \lambda_4 S_t^4}_{S_t} + \varepsilon_t,$$

où τ_t est la tendance (linéaire) et S_t est la composante saisonnière.

Le modèle s'écrit alors,

$$Y = X\theta + \varepsilon,$$

qui peut être exprimé sous forme matricielle,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \end{pmatrix} = \begin{pmatrix} 1 & 1 & S_{1,1} & S_{1,2} & S_{1,3} & S_{1,4} \\ 1 & 2 & S_{2,1} & S_{2,2} & S_{2,3} & S_{2,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t & S_{t,1} & S_{t,2} & S_{t,3} & S_{t,4} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_4 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_t \end{pmatrix}$$

où

— Y est le vecteur colonne des observations de la variable dépendante Y_t,

— X est la matrice des variables explicatives connue,

— θ est le vecteur des paramètres inconnus du modèle,

— ε est le vecteur des erreurs aléatoires.

Soit $Y = X\theta + \varepsilon$, l'écriture de l'estimateur des moindres carrés ordinaires s'écrit

$$\hat{\theta} = (X'X)^{-1}X'Y,$$

Toutefois, cette écriture n'est possible que si $X'X$ est inversible. ce qui n'est pas le cas lorsque certaines colonnes de X sont linéairement dépendantes, $X'X$ devient non inversible et l'estimateur des moindres carrés ordinaires (MCO) ne peut pas être directement calculé sous cette forme. Deux méthodes sont alors possibles pour faire malgré tout l'identification du modèle.

- Ne pas tenir compte de la constante, et identifier le modèle

$$Y_t = \beta_1 t + \delta_1 S_t^1 + \delta_2 S_t^2 + \delta_3 S_t^3 + \delta_4 S_t^4 + \varepsilon_t. \tag{2.1}$$

- Rajouter une contrainte, et identifier le modèle

$$\begin{cases} Y_t = \beta_0 + \beta_1 t + \lambda_1 S_t^1 + \lambda_2 S_t^2 + \lambda_3 S_t^3 + \lambda_4 S_t^4 + \varepsilon_t, \\ \text{sous contrainte :} \quad \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 0. \end{cases} \tag{2.2}$$

Cette dernière contrainte est arbitraire, mais correspond à une interprétation bien précise.

Exemple 2.2.4 *Nous considérons ici une série chronologique, mensuelle, comportant une forte saisonalité : trafic voyageur de la SNCF.*

TABLE 2.1 – Trafic voyageurs SNCF – Données mensuelles.

Année	JAN	FÉV	MAR	AVR	MAI	JUI	JUIL	AOÛ	SEP	OCT	NOV	DÉC
1963	1750	1560	1820	2090	1910	2410	3140	2850	2090	1850	1630	2420
1964	1710	1600	1800	2120	2100	2460	3200	2960	2190	1870	1770	2270
1965	1670	1640	1770	2190	2020	2610	3190	2860	2140	1870	1760	2360
1966	1810	1640	1860	1990	2110	2500	3030	2900	2160	1940	1750	2330
1967	1850	1590	1880	2210	2110	2480	2880	2670	2100	1920	1670	2520
1968	1834	1792	1860	2138	2115	2485	2581	2639	2038	1936	1784	2391
1969	1798	1850	1981	2085	2120	2491	2834	2725	1932	2085	1856	2553
1970	1854	1823	2005	2418	2219	2722	2912	2771	2153	2136	1910	2537
1971	2008	1835	2120	2304	2264	2175	2928	2738	2178	2137	2009	2546
1972	2084	2034	2152	2522	2318	2684	2971	2759	2267	2152	1978	2723
1973	2081	2112	2279	2661	2281	2929	3089	2803	2296	2210	2135	2862
1974	2223	2248	2421	2710	2505	3021	3327	3044	2607	2525	2160	2876
1975	2481	2428	2596	2923	2795	3287	3598	3118	2875	2754	2588	3266
1976	2667	2668	2804	2806	2976	3430	3705	3053	2764	2802	2707	3307
1977	2706	2586	2796	2978	3053	3463	3649	3095	2839	2966	2863	3375
1978	2820	2857	3306	3333	3141	3512	3744	3179	2984	2950	2896	3611
1979	3313	2644	2872	3267	3391	3682	3937	3284	2849	3085	3043	3541
1980	2848	2913	3248	3250	3375	3640	3771	3259	3206	3269	3181	4008

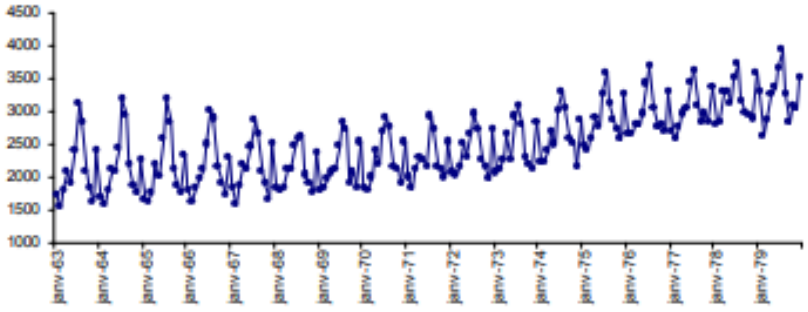


FIGURE 2.5 – Trafic voyageurs SNCF – Données mensuelles.

Supposons que les données commencent au 1er trimestre. Le modèle s'écrit alors, pour l'exemple du trafic SNCF :

$$\begin{pmatrix}5130\\6410\\8080\\5900\\5110\\6680\\8350\\5910\\5080\\\vdots\\Y_t\end{pmatrix} = \beta_0 \begin{pmatrix}1\\1\\1\\1\\1\\1\\1\\1\\1\\\vdots\\1\end{pmatrix} + \beta_1 \begin{pmatrix}1\\2\\3\\4\\5\\6\\7\\8\\9\\\vdots\\t\end{pmatrix} + \lambda_1 \begin{pmatrix}1\\0\\0\\0\\1\\0\\0\\0\\1\\\vdots\\S_t^1\end{pmatrix} + \lambda_2 \begin{pmatrix}0\\1\\0\\0\\0\\1\\0\\0\\0\\\vdots\\S_t^2\end{pmatrix} + \lambda_3 \begin{pmatrix}0\\0\\1\\0\\0\\0\\1\\0\\0\\\vdots\\S_t^3\end{pmatrix} + \lambda_4 \begin{pmatrix}0\\0\\0\\1\\0\\0\\0\\1\\0\\\vdots\\S_t^4\end{pmatrix} + \begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\\\varepsilon_6\\\varepsilon_7\\\varepsilon_8\\\varepsilon_9\\\vdots\\\varepsilon_t\end{pmatrix},$$

qui peut se réécrire, de façon matricielle,

$$\begin{pmatrix}5130\\6410\\8080\\5900\\5110\\6680\\8350\\5910\\5080\\\vdots\\Y_t\end{pmatrix} = \begin{pmatrix}1&1&1&0&0&0&0\\1&2&0&1&0&0&0\\1&3&0&0&1&0&0\\1&4&0&0&0&1&0\\1&5&1&0&0&0&0\\1&6&0&1&0&0&0\\1&7&0&0&1&0&0\\1&8&0&0&0&1&0\\1&9&1&0&0&0&0\\\vdots&\vdots&\vdots&\vdots&\vdots&\vdots&\vdots\\1&t&S_t^1&S_t^2&S_t^3&S_t^4&\cdots\end{pmatrix} \begin{pmatrix}\beta_0\\\beta_1\\\lambda_1\\\lambda_2\\\lambda_3\\\lambda_4\\\vdots\\\lambda_t\end{pmatrix} + \begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\varepsilon_3\\\varepsilon_4\\\varepsilon_5\\\varepsilon_6\\\varepsilon_7\\\varepsilon_8\\\varepsilon_9\\\vdots\\\varepsilon_t\end{pmatrix}.$$

Remarque 2.2.4 *Dans cet exemple, la matrice $X'X$ n'est pas inversible en raison d'une dépendance linéaire entre les colonnes de X. En effet, la première colonne (correspondant à la constante) est égale à la somme des quatre dernières (les composantes trimestrielles). Par conséquent, il n'est pas possible d'appliquer directement la méthode des moindres carrés ordinaires. Il devient alors nécessaire de recourir à l'une des deux méthodes alternatives permettant d'identifier le modèle malgré cette singularité.*

2. Estimateur des moindres carrés ordinaires (MCO) :

Solutions générales :

On considère un modèle de la forme :

$$Y_t = \sum_{i=1}^{m} \tau_t^i \beta_i + \sum_{j=1}^{p} S_t^j \lambda_j + \varepsilon_t, \quad \forall t = 1, \ldots, L.$$

La méthode des moindres carrés ordinaires (MCO) consiste à choisir les β_i et λ_j de façon à minimiser la somme des carrés des erreurs :

$$\left(\hat{\beta}, \hat{\lambda}\right) = \arg\min \sum_{t=1}^{L} \varepsilon_t^2 = \arg\min \sum_{t=1}^{L} \left[Y_t - \sum_{i=1}^{m} \tau_t^i \beta_i - \sum_{j=1}^{p} S_t^j \lambda_j \right]^2,$$

Notations : $\beta = (\beta_1, \ldots, \beta_m)'$, $\lambda = (\lambda_1, \ldots, \lambda_p)'$.

$$\tau = \begin{bmatrix} | & & | \\ \tau^1 & \cdots & \tau^m \\ | & & | \end{bmatrix} = \left[\tau_t^i\right]_{\substack{i=1,\ldots,m \\ t=1,\ldots,L}}, \quad \text{et} \quad S = \begin{bmatrix} | & & | \\ S^1 & \cdots & S^p \\ | & & | \end{bmatrix} = \left[S_t^j\right]_{\substack{j=1,\ldots,p \\ t=1,\ldots,L}}.$$

Le modèle s'écrit alors :

$$Y = \tau\beta + S\lambda + \varepsilon = \begin{bmatrix} \tau & S \end{bmatrix} \begin{bmatrix} \beta \\ \lambda \end{bmatrix} + \varepsilon = X\theta + \varepsilon.$$

et $\hat{\theta} = (\hat{\beta}; \hat{\lambda})'$ vérifie alors l'équation :

$$X'X\hat{\theta} = X'Y, \quad \text{soit} \quad \begin{bmatrix} \tau & S \end{bmatrix} \begin{bmatrix} \tau' \\ S' \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \tau'Y \\ S'Y \end{bmatrix},$$

et donc :

$$\begin{bmatrix} \hat{\beta} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} \tau'\tau & \tau'S \\ S'\tau & S'S \end{bmatrix}^{-1} \begin{bmatrix} \tau'Y \\ S'Y \end{bmatrix},$$

Ce qui donne les coefficients :

$$\hat{\beta} = \left[\tau'\tau - \tau'S(S'S)^{-1}S'\tau\right]^{-1} \left[\tau'Y - \tau'S(S'S)^{-1}S'Y\right],$$
$$\hat{\lambda} = \left[S'S - S'\tau(\tau'\tau)^{-1}\tau'S\right]^{-1} \left[S'Y - S'\tau(\tau'\tau)^{-1}\tau'Y\right].$$

Remarque 2.2.5 *S'il n'y a pas d'effet saisonnier, $Y = \tau\beta + \varepsilon$, et on retrouve le modèle linéaire usuel, avec pour estimateur mco $\hat{\beta} = \left[\tau'\tau\right]^{-1}\tau'Y$.*

- **Cas particulier : le modèle trimestriel de Buys-Ballot**

  Pour le modèle :

$$Y_t = \beta_0 + \beta_1 t + S_t^1\lambda_1 + S_t^2\lambda_2 + S_t^3\lambda_3 + S_t^4\lambda_4 + \varepsilon_t,$$

  il est possible d'expliciter les différents coefficients. L'équation

$$\begin{cases} \min\limits_{\beta,\lambda} \sum\limits_{t=1}^{L} \left(Y_t - \beta_0 - \beta_1 t - \sum\limits_{j=1}^{4} S_t^j\lambda_j\right)^2, \\ \text{sous contrainte :} \quad \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 = 0. \end{cases}$$

  peut se réécrire :

$$\min_{\beta,\lambda} \sum_{t=1}^{L} \left(Y_t - \beta_1 t - \sum_{j=1}^{4} S_t^j\delta_j\right)^2, \quad \text{où} \quad \begin{cases} \beta_0 = \dfrac{\delta_1 + \delta_2 + \delta_3 + \delta_4}{4}, \\ \lambda_j = \delta_j - \beta_0. \end{cases}$$

  En notant N le nombre d'années entières $\left(N = \frac{L}{4}\right)$, on pose :

— $\tilde{y}_i$: moyenne des Y_t relatives à l'année i,

— y_j : moyenne des y_t relatives au trimestre j,

— y : moyenne de toutes les observations y_t.

On a alors les estimateurs suivants :

$$\hat{\beta}_1 = \frac{3\sum_{i=1}^{N} i\,\tilde{y}_i - \frac{N(N+1)}{2}\bar{y}}{N(N^2-1)},$$

$$\hat{\delta}_j = \bar{y}_j - [j + 2(N-1)]\,\hat{\beta}_1, \quad \forall j = 1,2,3,4.$$

d'où finallement :

$$\begin{cases} \hat{\beta}_0 = \dfrac{\hat{\delta}_1 + \hat{\delta}_2 + \hat{\delta}_3 + \hat{\delta}_4}{4}, \\ \hat{\lambda}_j = \hat{\delta}_j - \hat{\beta}_0 \quad \forall j = 1,2,3,4. \end{cases}$$

- **Généralisation des formules de Buys-Ballot (tendance linéaire) :**

Les relations obtenues dans le cas précédent peuvent en fait être généralisées dans le cas d'une périodicité p, et en notant (de la même façon que précédemment) N le nombre d'années entières. Le modèle s'écrit alors :

$$Y_t = \beta_0 + \beta_1 t + S_t^1\,\lambda_1 + S_t^2\,\lambda_2 + \cdots + S_t^p\,\lambda_p + \varepsilon_t.$$

L'équation :

$$\begin{cases} \min\limits_{\beta,\,\lambda} \sum\limits_{t=1}^{L} \left(Y_t - \beta_0 - \beta_1 t - \sum\limits_{j=1}^{p} S_t^j\,\lambda_j \right)^2, \\ \text{sous contrainte :} \quad \lambda_1 + \lambda_2 + \lambda_3 + \cdots + \lambda_p = 0. \end{cases}$$

admet alors pour solution, en notant :

$$\hat{\beta}_1 = \frac{12}{p} \cdot \frac{\sum_{i=1}^{N} i\,\tilde{y}_i - \frac{N(N+1)}{2}\,\bar{y}}{N(N^2-1)},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \frac{Np+1}{2},$$

$$\hat{\lambda}_j = \tilde{y}_j - \bar{y} - \hat{\beta}_1 \cdot \left(j - \frac{p+1}{2} \right).$$

3. Propriétés statistiques des estimateurs :

Sous l'hypothèse $\mathbb{E}(\varepsilon_t) = 0$, les estimateurs (MCO) sont sans biais :

$$\mathbb{E}(\hat{\beta}_i) = \beta_i \quad \text{et} \quad \mathbb{E}(\hat{\lambda}_j) = \lambda_j.$$

La variance des estimateurs peut être estimée par :

$$\widehat{\mathrm{Var}} \begin{pmatrix} \hat{\beta} \\ \hat{\lambda} \end{pmatrix} = \hat{\sigma}^2 \begin{bmatrix} \tau'\tau & \tau' S \\ S'\tau & S'S \end{bmatrix}^{-1}, \quad \text{où} \quad \hat{\sigma}^2 = \frac{1}{L-m-p} \sum_{t=1}^{L} \hat{\varepsilon}_t^2.$$

Ce qui permet d'obtenir des intervalles de confiance sur les estimateurs.

Remarque 2.2.6 *Ces propriétés restent valables même dans le cas simple avec une seule variable explicative.*

De plus, après avoir sélectionné les variables explicatives et ajusté un modèle de régression, il est nécessaire de tracer les résidus afin de vérifier la validité des hypothèses. Plusieurs tracés doivent être réalisés afin de vérifier différents aspects du modèle ajusté et les hypothèses sous-jacentes. Nous allons maintenant les aborder tour à tour.

2.3 Evaluation du modèle

2.3.1 Analyse de l'ACF des résidus

Avec des données de séries chronologiques, il est fort probable que la valeur d'une variable observée au cours de la période actuelle soit similaire à sa valeur au cours des périodes précédentes. Par conséquent, lors de l'ajustement d'un modèle de régression à des données de séries chronologiques, il est fréquent de constater une autocorrélation dans les résidus. Dans ce cas, le modèle estimé ne respecte pas l'hypothèse d'absence d'autocorrélation dans les erreurs, et nos prévisions peuvent être inefficaces. Les prévisions d'un modèle comportant des erreurs autocorrélées restent non biaisées et ne sont donc pas « erronées », mais leurs intervalles de prédiction sont généralement plus grands que nécessaire. Il est donc conseillé de toujours examiner un graphique ACF des résidus.

Ensuite, il est toujours judicieux de vérifier si les résidus sont normalement distribués. En traçant leur histogramme, cela n'est pas indispensable pour les prévisions, mais cela simplifie grandement le calcul des intervalles de prédiction.

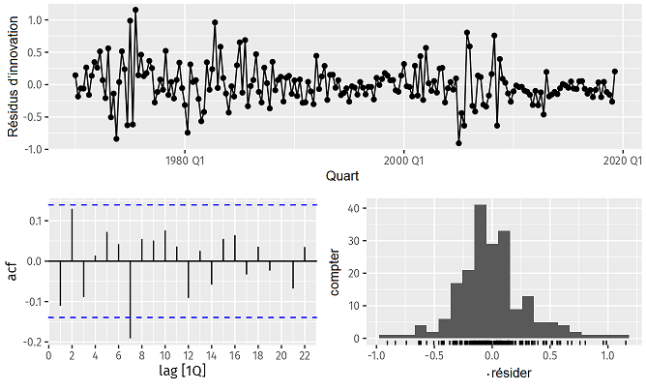


FIGURE 2.6 – Analyse des résidus d'un modèle de régression.

2.3.2 Analyse des résidus par rapport aux prédicteurs

On s'attend à ce que les résidus soient dispersés de manière aléatoire, sans présenter de schémas systématiques. Une manière simple et rapide de vérifier cela consiste à examiner les nuages de points des résidus par rapport à chacune des variables prédictives (explicatives). Si ces graphiques montrent un motif, cela peut indiquer que la relation est non linéaire et que le modèle devra être modifié en conséquence.

Il est également nécessaire de tracer les résidus en fonction des prédicteurs qui ne sont pas inclus dans le modèle. Si l'un de ces graphiques montre un motif, alors le prédicteur correspondant devrait peut-être ajouté au modèle, éventuellement sous une forme non linéaire.

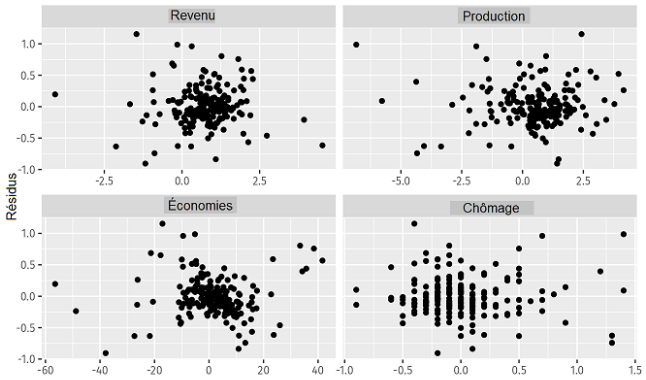


FIGURE 2.7 – Nuages de points des résidus par rapport à chaque prédicteur.

2.3.3 Analyse des résidus par rapport aux valeurs ajustées

Un graphique des résidus par rapport aux valeurs ajustées ne doit pas montrer de motif particulier. Si un motif est observé, cela peut indiquer la présence d'hétéroscédasticité dans les erreurs, ce qui signifie que la variance des résidus n'est peut-être pas constante. Si ce problème survient, il peut être nécessaire de transformer la variable à prédire (par exemple en utilisant le logarithme).

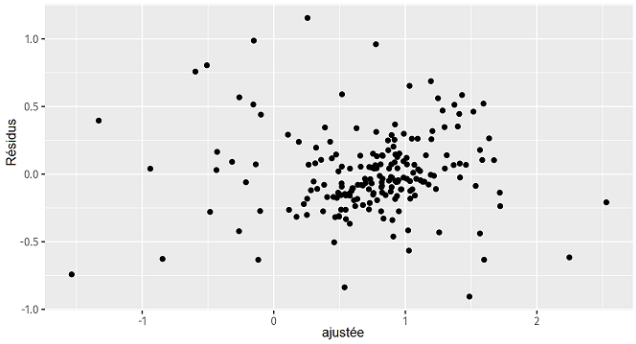


FIGURE 2.8 – Nuages de points des résidus par rapport aux valeurs ajustées.

Après avoir évalué le modèle à travers l'analyse des résidus, il est désormais nécessaire de passer à l'étape suivante : la validation, qui permet de juger de la qualité globale de l'ajustement et de la fiabilité des prévisions.

2.4 Validation du modèle

2.4.1 Coefficient de corrélation linéaire

Afin de confirmer qu'il est raisonnable d'approximer le nuage de points par une droite, on calcule le coefficient de corrélation linéaire r.

$$r = \frac{\mathrm{Cov}(t, y)}{\sigma_t \sigma_y}, \quad -1 \leq r \leq 1.$$

On considère que la corrélation linéaire est :

• Si $r(t, y) = 0$, on dira alors que t et y sont linéairement indépendants, l'éloignement des points du nuage avec la droite de régression de y en t est maximal.

• Si $r(t, y) = 1$, on peut donc parler de corrélation linéaire croissante totale positive, les points du nuage sont alors parfaitement alignés.

• Si $r(t, y) = -1$, on parle alors de corrélation linéaire décroissante totale négative, les points du nuage sont alors parfaitement alignés.

2.4.2 Analyse de la variance

• L'analyse de la variance est utile pour tester la significativité globale du modèle de régression, c'est-à-dire si l'ensemble des variables explicatives a une influence sur la variable à expliquer.

Afin de définir le *coefficient de détermination* qui permet de mesurer la qualité de l'ajustement linéaire, il convient d'analyser la décomposition de la variance totale de la variable dépendante. La variance totale s'exprime comme la somme de la variation expliquée par la régression et de la variation résiduelle. En effet :

$$SCT = SCR + SCE,$$

où SCT, SCR et SCE désignent les sommes de carrés suivantes :

$$SCT = \sum_{t=1}^{L} (y_t - \bar{y})^2,$$

$$SCE = \sum_{t=1}^{L} (\hat{y}_t - \bar{y})^2,$$

$$SCR = \sum_{t=1}^{L} (y_t - \hat{y}_t)^2.$$

Ensuite, l'analyse de la variance dépend du calcul des carrés moyens associés aux sommes SCR et SCE, obtenus en divisant les sommes de carrés par les nombres de degrés de liberté respectifs. Donc :

$$CMR = \frac{SCR}{L - k - 1}, \quad \text{et} \quad CME = \frac{SCE}{k}, \quad \forall k \in \mathbb{N}^*.$$

où k est le nombre total de paramètres estimés.

• Le coefficient de détermination est une interprétation du pourcentage de variation expliquée par la régression. Il se calcule comme suit :

$$R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}}, \quad 0 \leq R^2 \leq 1.$$

Cas particuliers :

— Si $R^2 = 0$, le modèle n'explique rien : les variables t et y ne sont pas corrélées linéairement.

— Si $R^2 = 1$, la relation linéaire explique toute la variation de y.

Le coefficient R^2 mesure la qualité de l'ajustement, mais il augmente automatiquement quand on ajoute des variables, même inutiles. Il ne permet donc pas de comparer objectivement plusieurs modèles entre eux.

• On définit donc un coefficient R^2 ajusté qui tient compte des degrés de liberté. Ce coefficient, noté $\bar{R}^2$, est défini comme suit :

$$\bar{R}^2 = 1 - \frac{\text{SCR}/(L-k-1)}{\text{SCT}/(L-1)} = 1 - \frac{(L-1)}{(L-k-1)}(1-R^2), \quad \forall k \in \mathbb{N}^*.$$

Remarque 2.4.1 *On a $\bar{R}^2$ est toujours inférieur à R^2, et ceci d'autant plus que le modèle contient un grand nombre de prédicteurs (variables explicatives).*

2.4.3 Test de signification globale du modèle

Pour vérifier la significativité globale du modèle de régression linéaire, on utilise le **test ANOVA**. Il permet de déterminer si au moins une variable explicative a un effet réel sur la variable dépendante.

Si le test est concluant, cela signifie que le modèle possède une capacité explicative globale.

Remarque 2.4.2 *Il existe une relation mathématique entre le R^2 et la statistique de test de signification globale (le F^* de Fisher) comme suit :*

$$F^* = \frac{(L-k-1)}{k}\frac{R^2}{1-R^2} \quad \forall k \in \mathbb{N}^*.$$

2.4.4 Test de contribution marginale

Pour vérifier si la contribution de chaque variable explicative est significative, on utilise le **test** t. Il permet de déterminer si chaque coefficient de régression est significativement différent de zéro.

Si le test est concluant, la variable explicative considérée a un effet significatif sur la variable dépendante.

Remarque 2.4.3 *Les détails techniques des tests sont présentés en Annexe A.*

Une fois la validation du modèle effectuée, il devient possible de procéder à la prévision de manière fiable.

2.5 Prévision avec Régression linéaire

Avant de mettre en œuvre un modèle de régression linéaire pour la prévision des séries chronologiques, il est important de considérer la question de la **stationnarité**. Bien que cette propriété ne soit pas une condition strictement requise pour appliquer une régression, son absence peut remettre en cause les hypothèses fondamentales du modèle, telles que l'indépendance et l'homoscédasticité des résidus. Elle peut également induire des relations trompeuses entre variables, appelées régressions fallacieuses. Afin d'éviter ces situations, il est souvent nécessaire de transformer les données (par différenciation ou centrage, par exemple) pour stabiliser leur structure. Enfin, lorsqu'on modélise une tendance ou une saisonnalité par une régression sur le temps, la stationnarité des résidus reste essentielle pour garantir la validité statistique du modèle et la fiabilité des prévisions.

2.5.1 Modèle de prévision pour tendance

Un des buts de la régression est de faire de la prévision, c'est-à-dire de prévoir la variable à expliquer y en présence d'une nouvelle valeur de la variable explicative t. Soit donc t_{L+h} une nouvelle valeur, pour laquelle nous voulons prédire y_{L+h}. Le modèle est toujours le même :

$$y_{L+h} = \beta_0 + \beta_1 t_{L+h} + \varepsilon_{L+h}, \quad \forall\, h \geq 1.$$

Ici, h représente l'horizon de prévision. Avec :

$$\mathbb{E}(\varepsilon_{L+h}) = 0 \quad , \quad \mathrm{Var}(\varepsilon_{L+h}) = \sigma^2, \quad \mathrm{Cov}(\varepsilon_t, \varepsilon_{L+h}) = 0 \quad , \quad \forall t = 1, \ldots, L.$$

Il est naturel de prédire la valeur correspondante via le modèle ajusté :

$$\hat{y}_{L+h} = \hat{\beta}_0 + \hat{\beta}_1 t_{L+h}.$$

Un intervalle de confiance pour cette prévision est de la forme :

— Pour β_0 :

$$\mathrm{IC}(\beta_0) = \left[\, \hat{\beta}_0 - St_{L-2}(1-\alpha/2) \cdot \hat{\sigma}_1 \,,\, \hat{\beta}_0 + St_{L-2}(1-\alpha/2) \cdot \hat{\sigma}_1 \,\right].$$

— Pour β_1 :

$$\mathrm{IC}(\beta_1) = \left[\, \hat{\beta}_1 - St_{L-2}(1-\alpha/2) \cdot \hat{\sigma}_2 \,,\, \hat{\beta}_1 + St_{L-2}(1-\alpha/2) \cdot \hat{\sigma}_2 \,\right].$$

où $St_{L-2}(1-\alpha/2)$ est le quantile de niveau $(1-\alpha/2)$ d'une loi de Student à $L-2$ degrés de liberté.

2.5.2 Modèle de prévision pour saisonalité

Soit $h \geq 1$. On suppose que le modèle reste valide en $L+h$, c'est à dire que :

$$Y_{L+h} = \sum_{i=1}^{m} \tau_{L+h}^{i}\, \beta_i + \sum_{j=1}^{p} S_{L+h}^{j}\, \lambda_j + \varepsilon_{L+h},$$

Avec $\mathbb{E}(\varepsilon_{L+h}) = 0, \quad \mathrm{Var}(\varepsilon_{L+h}) = \sigma^2, \quad$ et $\quad \mathrm{Cov}(\varepsilon_t, \varepsilon_{L+h}) = 0, \quad \forall t = 1, \ldots, L.$

La variable Y_{L+h} peut être approchée par :

$$\hat{Y}_L(h) = \sum_{i=1}^{m} \tau_{L+h}^{i}\, \hat{\beta}_i + \sum_{j=1}^{p} S_{L+h}^{j}\, \hat{\lambda}_j.$$

Cette prévision est la meilleure (au sens de l'erreur quadratique moyenne), linéaire en $Y_1, \ldots, Y_L$ et sans biais. Un intervalle de confiance pour cette prévision est de la forme :

$$\left[\hat{Y}_L(h) - St_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{\varepsilon}_h},\ \hat{Y}_L(h) + St_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{\varepsilon}_h}\right],$$

où $St_{1-\frac{\alpha}{2}}$ est le quantile d'ordre α de la loi de Student à $L-m-p$ degrés de liberté, et où

$$\begin{aligned}\hat{\varepsilon}_h = \widehat{\mathbb{E}}\left[\left(\hat{Y}_L(h) - Y_{L+h}\right)^2\right] &= \widehat{\mathrm{Var}}\left(\sum_{i=1}^{m} \tau_{L+h}^{i}\hat{\beta}_i + \sum_{j=1}^{p} S_{L+h}^{j}\hat{\lambda}_j - \varepsilon_{L+h}\right), \\ &= \begin{bmatrix}\hat{\beta}' & \hat{\lambda}'\end{bmatrix}\left[\widehat{\mathrm{Var}}\begin{pmatrix}\hat{\beta} \\ \hat{\lambda}\end{pmatrix}\right]\begin{bmatrix}\hat{\beta} \\ \hat{\lambda}\end{bmatrix} + \hat{\sigma}^2.\end{aligned}$$

2.5.3 Exemples instructifs :

1. Série avec tendance :

Exemple 2.5.1 *Les données suivantes représentent le nombre de champs découverts au cours des années de 1991 à 2000.*

TABLE 2.2 – Nombre de champs découverts par année.

Année (t)	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
y	62	63	67	69	70	75	79	82	84	86

Dans ce qui suit, nous allons identifier le type de tendance générale et estimer l'équation de la tendance linéaire correspondante, puis utiliser ce modèle pour prévoir le nombre de champs découverts en 2002.

1. Détermination du type de tendance générale :

En observant les valeurs de y, on remarque qu'elles augmentent au fil du temps. Donc, la tendance générale est croissante.

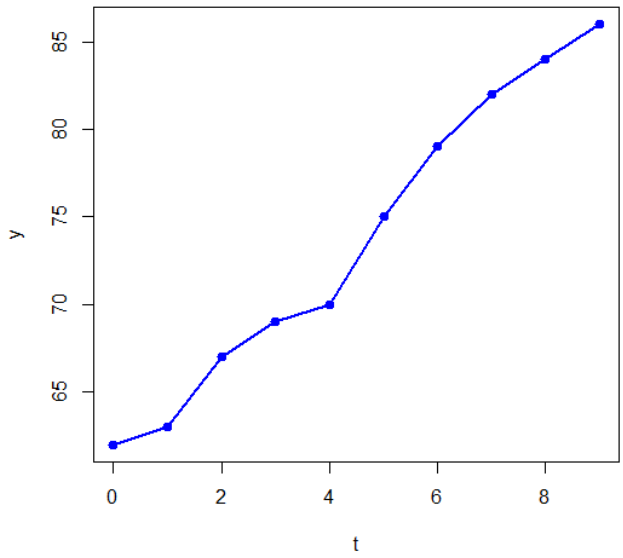


FIGURE 2.9 – Evolution de la série chronologique.

2. Estimation de l'équation de la tendance :

Le tableau suivant regroupe les données nécessaires à l'estimation de la tendance :

TABLE 2.3 – Tableau récapitulatif des calculs intermédiaires pour la régression.

Année	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
t	0	1	2	3	4	5	6	7	8	9
y	62	63	67	69	70	75	79	82	84	86
t^2	0	1	4	9	16	25	36	49	64	81
ty	0	63	134	207	280	375	474	574	672	774

Soient les valeurs moyennes et les sommes suivantes :

$$\bar{t} = 4,5 \quad ; \quad \bar{y} = 73,7 \quad ; \quad \overline{t^2} = 28,5 \quad ; \quad \overline{ty} = 355,3.$$

ainsi que la covariance et la variance :

$$\mathrm{Cov}(t,y) = 23,65 \quad ; \quad \mathrm{Var}(t) = 8,25.$$

Les coefficients de la droite de régression linéaire, sont alors estimés par la méthode des moindres carrés selon les formules :

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(t,y)}{\mathrm{Var}(t)} = \frac{23,65}{8,25} \approx 2,87 \quad ;$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{t} = 73,7 - 2,87 \times 4,5 \approx 60,79.$$

Ansi, l'équation de la tendance linéaire est :

$$\hat{y} = 60,79 + 2,87t.$$

3. Validation du modèle :

Pour analyser la relation entre les variables t et y, il convient d'abord de tracer le nuage de points, puis de calculer le coefficient de corrélation linéaire $r(t,y)$. Ce n'est que lorsque cette corrélation est suffisamment forte que l'on cherchera à déterminer la droite de régression de y en fonction de t.

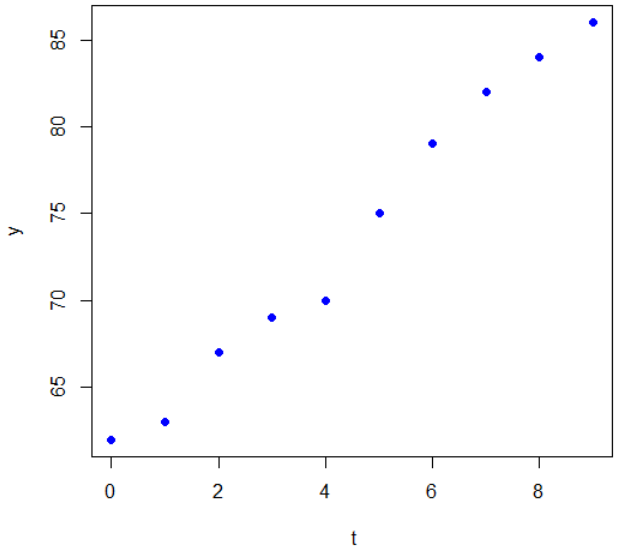


FIGURE 2.10 – Nuage de point.

Soit r le coefficient de corrélation linéaire défini par la relation :

$$r = \frac{\mathrm{Cov}(t,y)}{\sigma_t \sigma_y}, \quad -1 \leq r \leq 1.$$

On rappelle :

$$\mathrm{Var}(y) = 68,81 \quad \Rightarrow \quad \sigma_y = \sqrt{68,81} \approx 8,30.$$

$$\mathrm{Var}(t) = 8,25 \quad \Rightarrow \quad \sigma_t = \sqrt{8,25} \approx 2,872.$$

En utilisant les valeurs estimées, on obtient :

$$r = \frac{\mathrm{Cov}(t,y)}{\sigma_t \sigma_y} \approx 0,992.$$

Ce résultat $r = 0,992$ indique une liaison linéaire positive relativement forte entre les variables t et y.

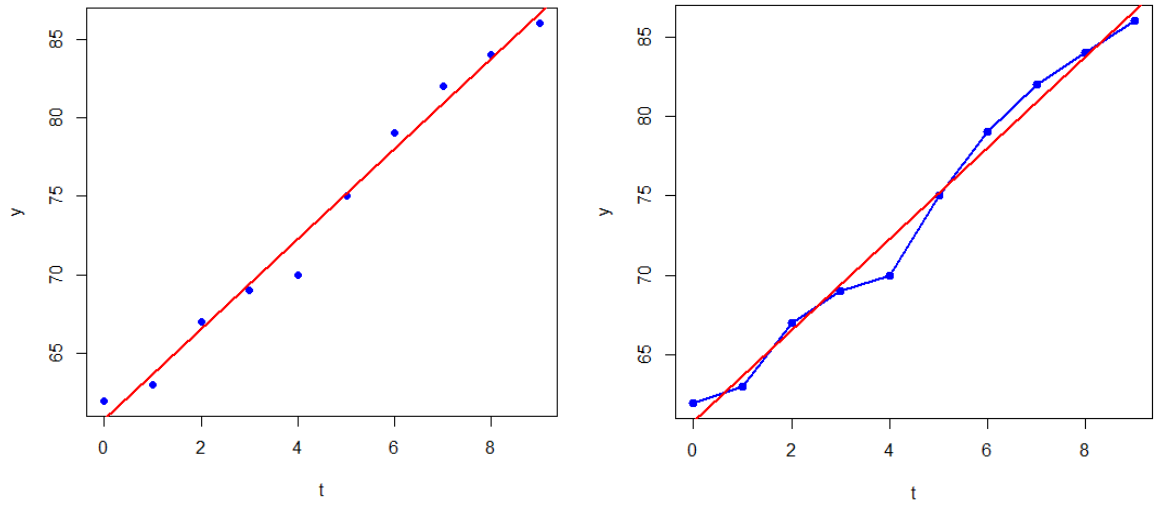


FIGURE 2.11 – Droite de la tendance.

• À partir du coefficient de corrélation r, on peut en déduire le coefficient de détermination R^2, qui permet d'évaluer la qualité de l'ajustement du modèle.

$$R^2 = r^2 \quad \Rightarrow \quad R^2 = \frac{(\,Cov(t,y))^2}{\sigma_t^2 \sigma_y^2} \approx 0,985.$$

Le coefficient de détermination $R^2 = 0,985$ montre que 98.5% de la variation de y est expliquée par t.

• Afin de préciser l'évaluation de la qualité du modèle, nous recourons désormais au coefficient de détermination ajusté $R^2_{\text{ajusté}}$, qui offre une estimation plus fiable.

$$R^2_{\text{ajusté}} = 1 - \left(\frac{(1-R^2)(L-1)}{L-k-1} \right) \approx 0,983.$$

Cela signifie que l'équation explique environ 98,3 % de la variance totale, ce qui indique que le modèle est très bien ajusté (ou très pertinent).

• Pour mieux comprendre les limites du modèle, il est important d'examiner les résidus, qui révèlent les écarts entre les valeurs observées et celles prédites.

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{1}{L-2}\sum_{i=1}^{L}(y_i - \hat{y}_i)^2} = \sqrt{\frac{(1-R^2)\hat{\sigma}_{yy}}{L-2}}.$$

$$\hat{\sigma}_{yy} = \sum_{i=1}^{L}(y_i - \bar{y})^2 = (L-1)\,\mathrm{Var}(y) \quad \Rightarrow \hat{\sigma}_\varepsilon = \sqrt{1,265} \approx 1,125.$$

• Bien que R^2 mesure l'ajustement global du modèle, il est nécessaire d'utiliser le test t pour vérifier si chaque variable y contribue significativement. Nous allons tester individuellement :

— si $\hat{\beta}_0$ (ordonnée à l'origine) est significatif,

— si $\hat{\beta}_1$ (pente) est significatif.

Calcul de l'erreur standard de la pente $\hat{\sigma}_{\hat{\beta}_1}$ et de l'ordonnée à l'origine $\hat{\sigma}_{\hat{\beta}_0}$:

$$\hat{\sigma}_{\hat{\beta}_1} = 0,1239 \quad ; \quad \text{et} \quad \hat{\sigma}_{\hat{\beta}_0} = 0,6615.$$

Calcul des statistiques t^* :

$$t^*_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \approx 23,14 \quad ; \quad et \quad t^*_{\hat{\beta}_0} = \frac{\hat{\beta}_0}{\hat{\sigma}_{\hat{\beta}_0}} \approx 91,91.$$

Le seuil critique pour un niveau de confiance de 95% avec 8 degrés de liberté :

$$St_{8,5\%} \approx 2,306$$

TABLE 2.4 – Résultats des tests de significativité du modèle de régression linéaire.

Test	**Statistique** t^*	**Seuil critique** $St_{8,5\%}$	**Décision**
Test de $\hat{\beta}_1$ (pente)	$t^*_{\hat{\beta}_1} \approx 23,14$	2,306	significatif
Test de $\hat{\beta}_0$ (constante)	$t^*_{\hat{\beta}_0} \approx 91,91$	2,306	significatif

• Après avoir évalué la significativité individuelle des variables grâce au test t. Nous allons tester la qualité globale du modèle par ANOVA, en comparant la variance expliquée à la variance résiduelle.

Dans ce modèle, nous avons $L = 10$ observations et un $r = 0{,}992$. Les degrés de liberté sont donc de 1 pour le modèle et de $L - 2 = 8$ pour l'erreur. En utilisant la formule statistique F^* :

$$F^* = \frac{R^2}{1 - R^2} \times (L - 2) = 535,24.$$

On va comparer F^* au seuil critique de la loi de Fisher $F_{\text{théorique}}(1, 8, \alpha)$.

Donc, pour un niveau de risque $\alpha = 5\%$, on a $F_{\text{théorique}}(1, 8, 5\%) \approx 5,32$.

Or :

$$F^* = 535.24 > F_{\text{théorique}} = 5,32.$$

Puisque $F^* > F_{\text{théorique}}$, on rejette l'hypothèse nulle H_0. La régression est hautement significative au seuil de 5% de risque.

4. Prévision pour l'année 2002 :

Pour calculer la valeur de $\hat{y}$ pour l'année 2002, nous devons utiliser l'équation du modèle de régression :

$$\hat{y}_{2002} = \hat{\beta}_0 + \hat{\beta}_1 t,$$

On connaît les valeurs suivantes :

$$\hat{\beta}_0 = 60{,}79 \quad ; \quad \hat{\beta}_1 = 2{,}87 \quad ; \quad t = 11.$$

Nous pouvons donc calculer $\hat{y}_{2002}$:

$$\hat{y}_{2002} = 60{,}79 + 2{,}87 \times 11 = 92,36 \approx 92.$$

Ainsi, la valeur de $\hat{y}$ pour l'année 2002 est $\boxed{92}$ champs.

2. Série agrégée par trimestre :

Exemple 2.5.2 *Consiédérons la série du trafic SNCF agrégée par trimestre, représentée ci-dessous, avec en ligne les années, et en colonne les trimestres.*

TABLE 2.5 – Valeurs observées des variables explicatives.

i/j	1	2	3	4	$\bar{y}_i$
1	5130	6410	8080	5900	6380
2	5110	6680	8350	5910	6513
3	5080	6820	8190	5990	6520
4	5310	6600	8090	6020	6505
5	5320	6800	7650	6110	6470
6	5486	6738	7258	6111	6398
7	5629	6696	7491	6494	6578
8	5682	7359	7836	6583	6865
9	5963	6743	7844	6692	6811
10	6270	7524	7997	6853	7161
11	6472	7871	8188	7207	7435
12	6892	8236	8978	7561	7917
13	7505	9005	9591	8608	8677
14	8139	9212	9522	8816	8922
15	8088	9494	9583	9204	9092
16	8983	9986	9907	9457	9583
17	8829	10340	10070	9669	9727
18	9009	10265	10236	10458	9992
$\bar{y}_j$	6605	7932	8603	7425	7641

représentée ci-dessous,

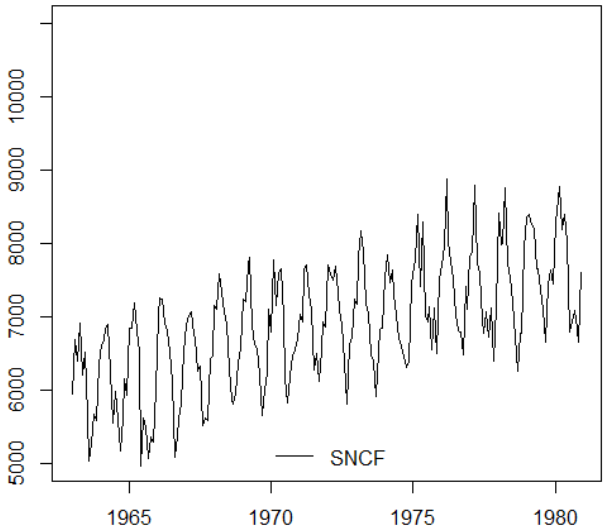


FIGURE 2.12 – Série du trafic SNCF agrégée par trimestre.

Considérons alors un modèle de la forme suivante, avec une saisonnalité en 4 composantes (les données étant trimestrielles : chaque composante correspondant à un trimestre), et une tendance supposée linéaire ($\tau_t = \beta_0 + \beta_1 t$) :

$$Y_t = \beta_0 + \beta_1 t + S_t^1\lambda_1 + S_t^2\lambda_2 + S_t^3\lambda_3 + S_t^4\lambda_4 + \varepsilon_t,$$

Compte tenu de la sur-identification de ce modèle, on rajoute la contrainte que la somme des λ_j soit nulle (c'est-à-dire que la composante saisonnière soit centrée : $\mathbb{E}(S_t) = 0$). On peut alors faire l'estimation de la façon suivante :

(i) On estime le modèle (2.1), c'est-à-dire sans contrainte, et sans constante β_0.

(ii) Et on se ramène au modèle (2.2) en utilisant les relations.

- **Calcul direct des estimateurs :**

Les calculs ont été fait ici sous les formules mathématiques classiques :

TABLE 2.6 – Tableau des composantes trimestrielles.

Année	i \ j	1	2	3	4	$\tilde{y}$	$i \times \tilde{y}_i$
1963	1	5130	6410	8080	5900	6380,00	6380,00
1964	2	5110	6680	8350	5910	6512,50	13025,00
1965	3	5080	6820	8190	5990	6520,00	19560,00
1966	4	5310	6600	8090	6020	6505,00	26020,00
1967	5	5320	6800	7650	6110	6470,00	32350,00
1968	6	5486	6738	7258	6111	6398,25	38389,50
1969	7	5629	6696	7491	6494	6577,50	46042,50
1970	8	5682	7359	7836	6583	6865,00	54920,00
1971	9	5963	6743	7844	6692	6810,50	61294,50
1972	10	6270	7524	7997	6853	7161,00	71610,00
1973	11	6472	7871	8188	7207	7434,50	81779,50
1974	12	6892	8236	8978	7561	7916,75	95001,00
1975	13	7505	9005	9591	8608	8677,25	112804,25
1976	14	8139	9212	9522	8816	8922,25	124911,50
1977	15	8088	9494	9583	9204	9092,25	136383,75
1978	16	8983	9986	9907	9457	9583,25	153332,00
1979	17	8829	10340	10070	9669	9727,00	165359,00
1980	18	9009	10265	10236	10458	9992,00	179856,00
	$\bar{y}_j$	6605	7932	8603	7425	7641, 39	

Pour chacune des années et chacun des trimestres, il est possible de calculer des moyennes. Ainsi, la moyenne pour l'année 1963 est de 6380, et celle pour l'année 1973 est de 7435. De façon analogue, la moyenne pour le premier trimestre est de 6605, et de 8603 pour le troisième trimestre. La moyenne totale est alors de 7641 pour ces 72 observations. Aussi, $N = 18$ (on a 18 années d'observations), et la pente de la droite de tendance est donnée par :

$$\hat{\beta}_1 = \frac{3}{N(N^2-1)}\left(\sum_{i=1}^{N} i\,\bar{y}_i - \frac{N(N+1)}{2}\,\bar{y}\right) = \frac{3}{18(18^2-1)}(1\,419\,019 - 1\,306\,678) = 57{,}97$$

En utilisant les moyennes par trimestre, et par année, données dans le tableau ci-dessus, et

$$\hat{\delta}_j = y_j - [j + 2(N-1)]\,\hat{\beta}_1, \quad \text{et donc} \quad \begin{cases} \hat{\delta}_1 = 6605 - 35 \times 57{,}97 \approx 4577, \\ \hat{\delta}_2 = 7932 - 36 \times 57{,}97 \approx 5845, \\ \hat{\delta}_3 = 8603 - 37 \times 57{,}97 \approx 6459, \\ \hat{\delta}_4 = 7425 - 38 \times 57{,}97 \approx 5222. \end{cases}$$

d'où finallement :

$$\begin{cases} \hat{\beta}_0 &= \frac{\hat{\delta}_1+\hat{\delta}_2+\hat{\delta}_3+\hat{\delta}_4}{4} \approx 5526, \\ \hat{\lambda}_j &= \hat{\delta}_j - \hat{\beta}_0. \end{cases} \quad \begin{cases} \hat{\lambda}_1 = 4577 - 5526 = -949, \\ \hat{\lambda}_2 = 5845 - 5526 = +320, \\ \hat{\lambda}_3 = 6459 - 5526 = +933, \\ \hat{\lambda}_4 = 5222 - 5526 = -304. \end{cases}$$

Aussi, le modèle s'écrit :

$$\hat{Y}_t = 5526 + 58t - 949S_t^1 + 320S_t^2 + 933S_t^3 - 304S_t^4.$$

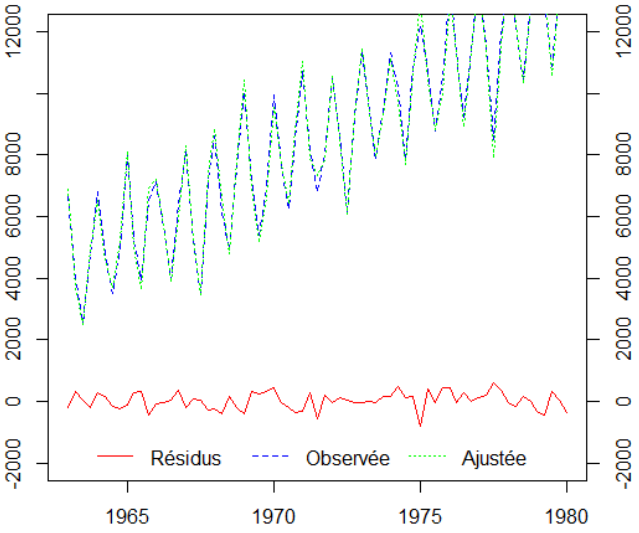


FIGURE 2.13 – Graphiques diagnostiques du modèle de régression.

La série ajustée (ci-dessous à gauche) correspond à la série :

$$\hat{Y}_t = Y_t - \varepsilon_t = \sum_{i=1}^{m} \tau_t^i \beta_i + \sum_{j=1}^{p} S_t^j \lambda_j,$$

Avec (τ_t) en trait plein, et $\hat{Y}_t$ en pointillés. Cette série pourra être prolongée afin de faire de la prévision. La série Corrigées des Variations Saisonnières (CVS-ci-dessous à droite) correspond à la série :

$$\hat{X}_t = Y_t - \hat{S}_t = \sum_{i=1}^{m} \tau_t^i \beta_i + \varepsilon_t.$$

Remarque 2.5.1 *La composante saisonnière S_t correspond à $S_t = \sum_{j=1}^{p} S_t^j \lambda_j$. Elle vérifie alors $\mathbb{E}(S_t) = 0$. Cette propriété n'est pas vérifiée dans le modèle sans constante.*

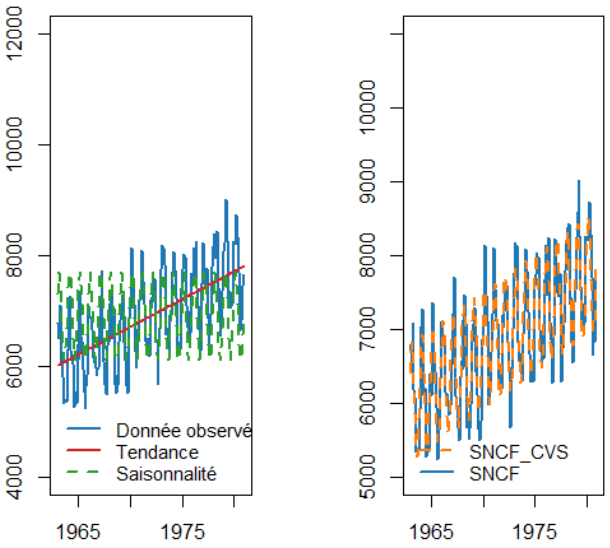


FIGURE 2.14 – Trajectoire d'une série ajustée et série corrigée trimestrielle.

3. Analyse sur données mensuelles :

La méthode décrite ci-dessus donne les résultats suivants :

TABLE 2.7 – Tableau des composantes mensuelles.

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC	$\tilde{y}_i$
1963	1750	1560	1820	2090	1910	2410	3140	2850	2090	1850	1630	2420	2127
1964	1710	1600	1800	2120	2100	2460	3200	2960	2190	1870	1770	2270	2171
1965	1670	1640	1770	2190	2020	2610	3190	2860	2140	1870	1760	2360	2173
1966	1810	1640	1860	1990	2110	2500	3030	2900	2160	1940	1750	2330	2168
1967	1850	1590	1880	2210	2110	2480	2880	2670	2100	1920	1670	2520	2157
1968	1834	1792	1860	2138	2115	2485	2581	2639	2038	1936	1784	2391	2133
1969	1798	1850	1981	2085	2120	2491	2834	2725	1932	2085	1856	2553	2192
1970	1854	1823	2005	2418	2219	2722	2912	2771	2153	2136	1910	2537	2288
1971	2008	1835	2120	2304	2264	2175	2928	2738	2178	2137	2009	2546	2270
1972	2084	2034	2152	2522	2318	2684	2971	2759	2267	2152	1978	2723	2387
1973	2081	2112	2279	2661	2281	2929	3089	2803	2296	2210	2135	2862	2478
1974	2223	2248	2421	2710	2505	3021	3327	3044	2607	2525	2160	2876	2639
1975	2481	2428	2596	2923	2795	3287	3598	3118	2875	2754	2588	3266	2892
1976	2667	2668	2804	2806	2976	3430	3705	3053	2764	2802	2707	3307	2974
1977	2706	2586	2796	2978	3053	3463	3649	3095	2839	2966	2863	3375	3031
1978	2820	2857	3306	3333	3141	3512	3744	3179	2984	2950	2896	3611	3194
1979	3313	2644	2872	3267	3391	3682	3937	3284	2849	3085	3043	3541	3242
1980	2848	2913	3248	3250	3375	3640	3771	3259	3206	3269	3181	4008	3331
$\bar{y}_j$	2195	2101	2309	2555	2489	2888	3249	2928	2426	2359	2205	2861	2547

qui donne les coefficients suivants :

TABLE 2.8 – Tableau de calcul des estimateurs.

$\bar{\beta}_1$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\delta}_4$	$\hat{\delta}_5$	$\hat{\delta}_6$	$\hat{\delta}_7$	$\hat{\delta}_8$	$\hat{\delta}_9$	$\hat{\delta}_{10}$	$\hat{\delta}_{11}$	$\hat{\delta}_{12}$
9,82	1038	943	1156	1380	1293	1667	1938	1517	1135	1123	975	1618

Ce qui donne la série ajustée (à gauche) et la série corrigée des variations saisonnières (à droite).

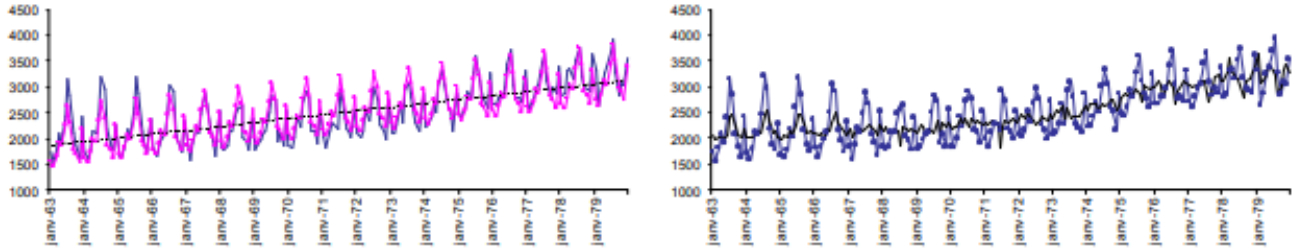


FIGURE 2.15 – Trajectoire d'une série ajustée et série corrigée mensuelle.

Dans l'exemple considéré précédemment, en données mensuelles, considérons désormais l'ensemble des données entre janvier 1970 et décembre 1980, et considérons le modèle suivant :

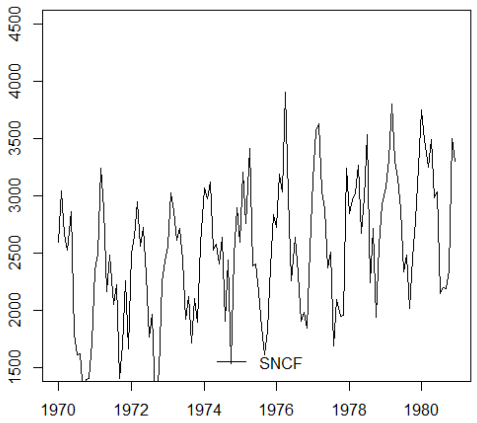
$$Y_t = \beta_1 t + S_t^1\delta_1 + S_t^2\delta_2 + S_t^3\delta_3 + S_t^4\delta_4 + S_t^5\delta_5 + S_t^6\delta_6 + S_t^7\delta_7 + S_t^8\delta_8 + S_t^9\delta_9 + S_t^{10}\delta_{10} + S_t^{11}\delta_{11} + S_t^{12}\delta_{12} + \varepsilon_t$$

L'estimation par la méthode des moindres carrés donne l'estimation suivante :

TABLE 2.9 – Tableau de calcul des estimateurs.

$\hat{\beta}_1$	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	$\hat{\lambda}_4$	$\hat{\lambda}_5$	$\hat{\lambda}_6$	$\hat{\lambda}_7$	$\hat{\lambda}_8$	$\hat{\lambda}_9$	$\hat{\lambda}_{10}$	$\hat{\lambda}_{11}$	$\hat{\lambda}_{12}$
9.82	1038	943	1156	1380	1293	1667	1938	1517	1135	1123	975	1618

avec les estimations d'écart-types suivantes :



LS // Dependent Variableis SNCF
Sample: 1970:01 1980:12
Included observations: 132 after adjusting endpoints

Variable	Coefficient	Std. Error	T-Statistic	Prob.
TEMPS	9.820391	0.316040	31.07324	0.0000
MOIS01	1038.316	61.85432	16.78647	0.0000
MOIS02	943.3138	62.08883	15.19297	0.0000
MOIS03	1156.312	62.32406	18.55321	0.0000
MOIS04	1380.400	62.55999	22.06522	0.0000
MOIS05	1292.943	62.79664	20.58937	0.0000
MOIS06	1667.396	63.03398	26.45233	0.0000
MOIS07	1938.121	63.27200	30.63157	0.0000
MOIS08	1516.664	63.51071	23.88045	0.0000
MOIS09	1135.480	63.75009	17.81143	0.0000
MOIS10	1122.751	63.99014	17.54568	0.0000
MOIS11	975.1121	64.23084	15.18137	0.0000
MOIS12	1618.201	64.47220	25.09920	0.0000

R-squared	0.932001	Mean dependent var	2793.386
Adjusted R-squared	0.925144	S.D. dependent var	503.6128
S.E. of regression	137.7877	Akaike info criterion	9.944720
Sum squared resid	2259269.	Schwartz criterion	10.22863
Log likelihood	-830.6514	F-statistic	135.9186
Durbin-Watson stat	1.284049	Prob(F-statistic)	0.000000

FIGURE 2.16 – Modélisation du trafic SNCF à l'aide d'un ajustement linéaire.

Comme le montre la sortie ci-dessus à droite, tous les paramètres sont significatifs, le R^2 est relativement bon (93 %), et la statistique de Fisher est suffisamment grande pour valider le modèle. La courbe de gauche ci-dessous correspond à la prévision du nombre de voyageurs pour les années 1982 et 1983, et l'intervalle de confiance de cette prévision est donné à droite.

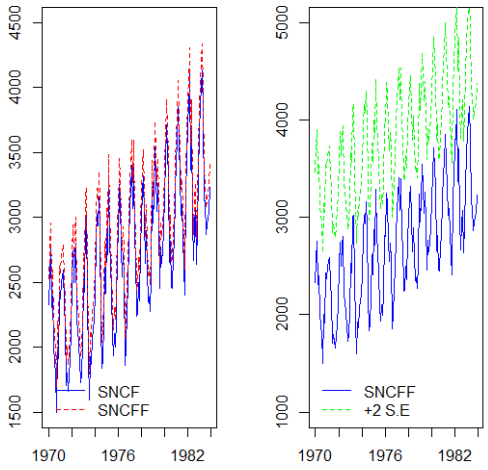


FIGURE 2.17 – Prévision du SNCF pour 1982 et 1983 et l'IC.

À 95 %, l'intervalle de confiance correspond à la prévision ±145 (soit une prévision à ±5 %). Si cette prévision est aussi robuste, c'est aussi parce que l'on a restreint l'intervalle d'étude à la période 1970–1980, en supprimant les premières années. Les résidus ainsi obtenus sont représentés ci-dessous.

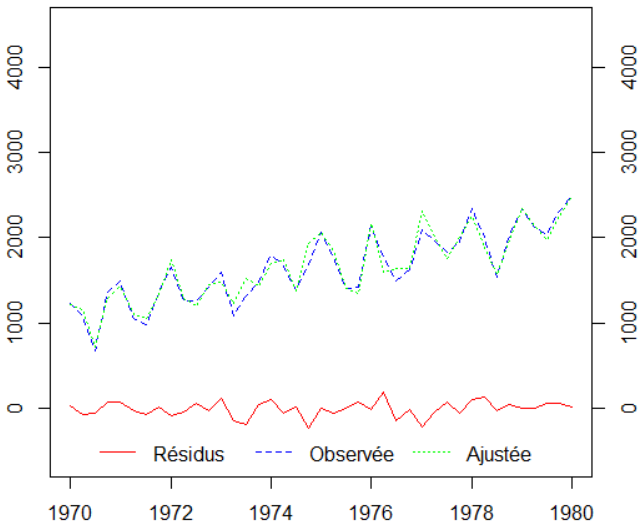


FIGURE 2.18 – Graphiques des résidus du modèle de régression.

2.5.4 Limites et avantages de la régression linéaire

À l'issue de ce cadre théorique consacré à la régression linéaire appliquée aux séries chronologiques, plusieurs avantages et limites peuvent être dégagés. La régression linéaire, qu'elle soit simple, multiple ou temporelle, offre une approche claire, mathématiquement rigoureuse et relativement simple à mettre en œuvre. Elle permet de modéliser des tendances et des effets saisonniers de manière directe, avec des coefficients faciles à interpréter. Cette méthode s'adapte bien aux séries présentant une structure linéaire, et elle est accompagnée de nombreux outils d'évaluation (résidus, ACF, R^2, tests de significativité) qui permettent de valider le modèle et d'estimer la qualité des prévisions.

Cependant, la régression linéaire repose sur des hypothèses fortes linéarité, indépendance et homoscédasticité des erreurs qui sont souvent difficiles à vérifier dans le cadre des séries temporelles. L'autocorrélation des résidus, la présence de valeurs extrêmes, la multicolinéarité entre prédicteurs ou encore l'existence de relations non linéaires peuvent rendre les estimations instables et les prévisions peu fiables. Par ailleurs, les composantes complexes comme la saisonnalité irrégulière ou les effets de changement de structure ne sont pas toujours bien capturés par un modèle linéaire classique.

Ainsi, si la régression linéaire constitue un point de départ pertinent pour modéliser et prévoir certaines séries, elle doit être utilisée avec précaution, et parfois complétée ou remplacée par des modèles plus adaptés à la dynamique temporelle réelle des données.

Chapitre 3

Application pratique de la régression linéaire avec R

Dans ce chapitre, nous mettons en pratique les méthodes de régression linéaire à l'aide du logiciel R (version 4.4.2), en mobilisant des *packages* spécifiques , notamment : `readr`, `readxl`, `dplyr`, `tidyr`, `ggplot2`, `stats`, `broom`, `car`, `performance`, ainsi que `forecast` ou `modelr` pour certaines fonctions de prévision. À partir de données réelles, nous illustrons les différentes étapes de l'analyse : importation, exploration, ajustement du modèle, évaluation et prévision. Cette mise en œuvre s'appuie sur des références théoriques solides et des exemples appliqués tirés de la littérature . [1][2] [5] [10] [11] [13] [14] et [16] .

3.1 Étude de la série chronologique par RLS

3.1.1 Présentation des données

Les données utilisées dans cette étude correspondent à la consommation annuelle d'électricité, exprimée en gigawattheures (GWh), pour l'Hôpital Dr. Benzerdjeb de la wilaya d'Ain Témouchent sur la période allant de 2014 à 2023. Ces données, issues des archives internes des services locaux d'agence commerciale de Sonelgaz, sont présentées dans le tableau ci-dessous :

Table 3.1 – Consommation d'électricité à l'Hôpital Dr. Benzerdjeb.

Année	Consommation (GWh)
2014	5 578 320
2015	5 812 870
2016	6 492 630
2017	6 978 410
2018	7 437 980
2019	7 562 500
2020	7 935 210
2021	8 628 340
2022	9 193 120
2023	9 752 880

3.1.2 Visualisation des données

Pour explorer visuellement la relation entre les années et la consommation en GWh, nous utilisons la commande `plot(donnees$Annee, donnees$Consommation_GWh)` afin de tracer la série chronologique et détecter d'éventuelles tendances ou variations.

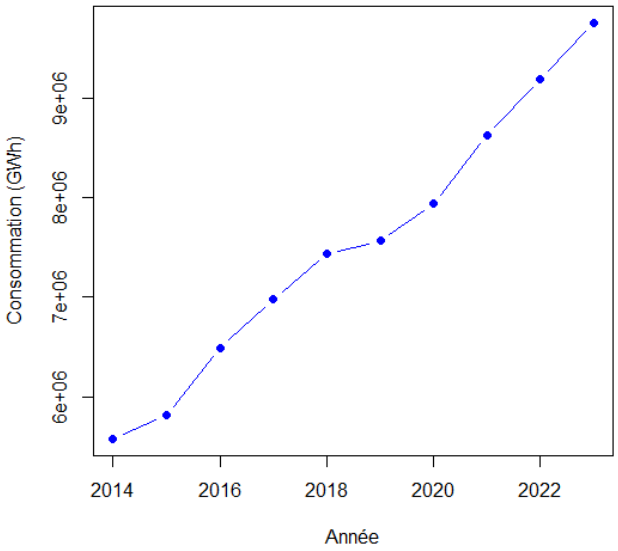


FIGURE 3.1 – Évolution de la consommation d'électricité (Hôpital Dr Benzerdjeb).

Les données présentées ci-dessus montrent l'évolution annuelle de la consommation d'électricité entre 2014 et 2023. On observe une tendance générale à la hausse.

3.1.3 Ajustement du modèle de régression linéaire

Pour expliquer la consommation anuelle d'électricité, nous ajustons un modèle linéaire de la forme :

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t,$$

où y_t représente la consommation en GWh à l'année t, β_0 l'ordonnée à l'origine, β_1 le coefficient de tendance, et ε_t l'erreur aléatoire.

L'ajustement du modèle est réalisé dans `R` à l'aide de la commande :

`lm(Consommation_GWh ~ Annee, data = donnees).`

```
>  modele

Call:
lm(formula = Consommation_GWh ~ Annee, data = donnees)

Coefficients:
(Intercept)        Annee
 -909269999       453977
```

FIGURE 3.2 – Ajustement du modèle sur les données observées.

Le modèle ajusté est de la forme :

$$\hat{y}_t = -909\,270 + 453\,980t.$$

Pour visualiser la droite de régression ajustée au modèle estimé, on utilise la commande `abline(modele)`.

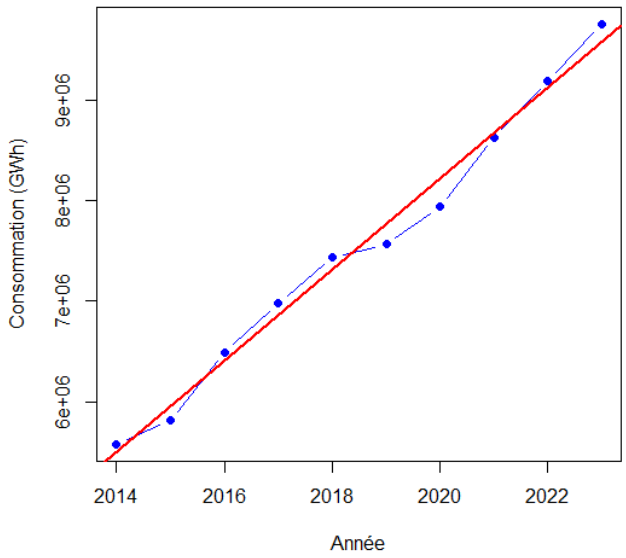


FIGURE 3.3 – La droite de l'équation d'une tendance.

3.1.4 Évaluation du modèle

Test d'indépendance des résidus (ACF) :

Pour vérifier l'indépendance des résidus, nous utilisons la commande `acf(residuals(modele))` dans `R`, qui permet de tracer la fonction d'autocorrélation des résidus. Cette étape est essentielle pour détecter d'éventuelles dépendances temporelles non prises en compte par le modèle.

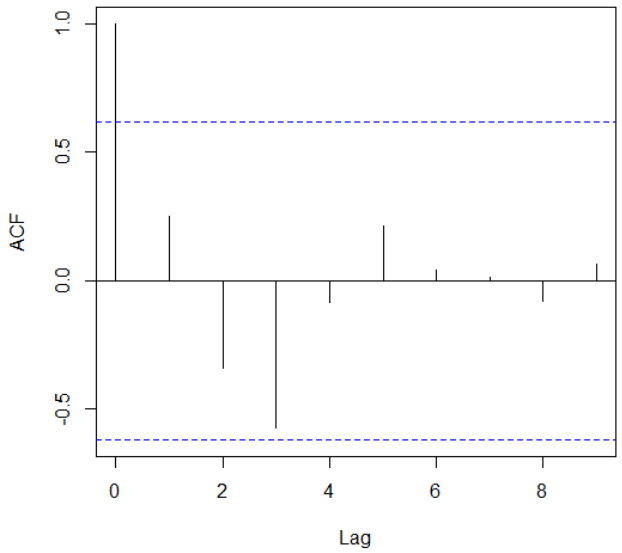


FIGURE 3.4 – ACF des résidus.

Le graphique ACF des résidus montre que toutes les barres (autres que celle à décalage = 0) sont comprises à l'intérieur des bandes bleues de confiance à 95 %. Cela signifie que les valeurs d'autocorrélation aux différents décalages (lags) ne sont pas statistiquement significatives. Donc, on ne détecte pas d'autocorrélation dans les résidus.

Test des résidus par rapport aux prédicteur :

Pour examiner la relation entre les résidus et la variable explicative, nous traçons le nuage de points à l'aide de la commande : `plot(donnees$Annee, residuals(modele))` dans R. Ce graphique permet de vérifier visuellement l'absence de structure particulière, condition essentielle pour valider l'hypothèse d'indépendance entre les résidus et le prédicteur.

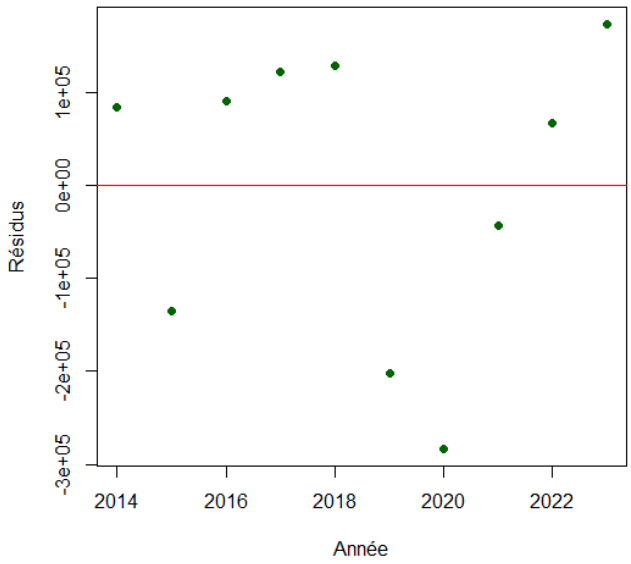


FIGURE 3.5 – Résidus en fonction de l'année.

Le graphique des résidus révèle une certaine variabilité au fil des années. Cela peut indiquer que le modèle tend à surestimer ou sous-estimer légèrement la consommation selon les périodes. La présence de légers motifs dans les résidus suggère qu'il pourrait exister des aspects de la dynamique des données non entièrement capturés par le modèle actuel.

Cependant, malgré ces observations, le modèle demeure globalement pertinent et performant, car il parvient à capter la tendance principale de la consommation d'électricité sur la période étudiée.

Test des résidus par rapport aux valeur ajustée :

Pour évaluer la qualité de l'ajustement du modèle, nous réalisons un graphique des résidus en fonction des valeurs ajustées à l'aide de la commande : `plot(fitted(modele), residuals(modele))`.

Ce diagnostic visuel permet de vérifier l'homoscédasticité : variance constante des résidus, et de détecter d'éventuelles structures non capturées par le modèle.

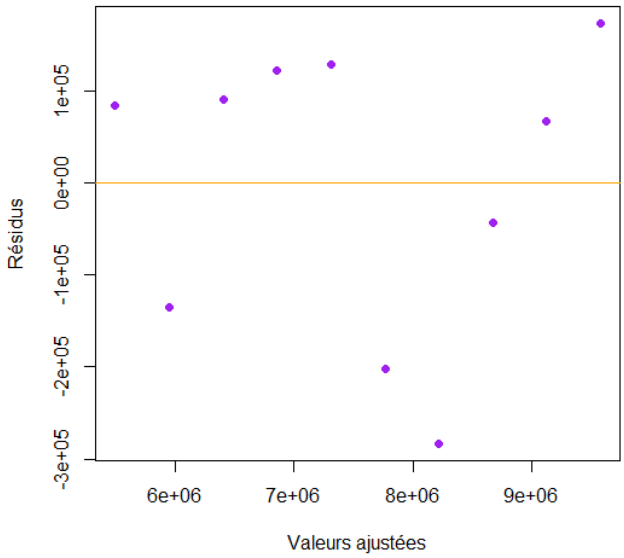


FIGURE 3.6 – Résidus en fonction de valeur ajustée.

Le graphique montre que les résidus sont globalement bien répartis, sans motif très structuré, ce qui suggère que le modèle capte une bonne partie de la relation entre les variables. Cela témoigne d'un ajustement globalement satisfaisant.

Test de normalité des résidus :

La normalité des résidus est vérifiée à l'aide du test de Shapiro-Wilk, en utilisant la commande : `shapiro.test(residuals(modele))`. Ce test permet d'évaluer si les résidus suivent une distribution normale, hypothèse fondamentale pour la validité des tests statistiques associés au modèle de régression.

```
> shapiro.test(residuals(modele))

        Shapiro-Wilk normality test

data:  residuals(modele)
W = 0.8816, p-value = 0.1361
```

FIGURE 3.7 – Test de normalité des résidus.

Le test de normalité de Shapiro-Wilk sur les résidus donne $W = 0.8816$ et une p-value $= 0.1361$. Comme $p > 0.05$, on ne rejette pas l'hypothèse nulle H_0 : les résidus suivent une loi normale.

3.1.5 Validation de modèle

La validation globale du modèle repose sur la **statistique de Fisher (F-statistic)**. Celle-ci est comparée à la valeur critique F_{tab} extraite de la loi de Fisher avec $k-1$ et $L-k$ degrés de liberté, où k est le nombre de paramètres estimés et L la taille de l'échantillon.

L'hypothèse nulle H_0, selon laquelle le modèle n'apporte aucune amélioration significative par rapport à un modèle sans prédicteurs, est rejetée si $F^* > F_{\text{tab}}$.

Cette validation est effectuée à l'aide de la commande `summary(modele)` dans R.

```
> summary(modele)

Call:
lm(formula = Consommation_GWh ~ Annee, data = donnees)

Residuals:
    Min      1Q  Median      3Q     Max
-282982 -112534   75483  114199  172756

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -909269999   37022885  -24.56 8.07e-09 ***
Annee           453977      18333   24.76 7.56e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 166500 on 8 degrees of freedom
Multiple R-squared:  0.9871,    Adjusted R-squared:  0.9855
F-statistic: 613.2 on 1 and 8 DF,  p-value: 7.559e-09
```

FIGURE 3.8 – Résumé statistique du modèle de régression linéaire simple.

On remarque que les valeurs minimales et maximales des résidus indiquent l'absence d'erreurs particulièrement extrêmes, ce qui suggère un bon ajustement global du modèle. Par ailleurs, le coefficient de détermination $R^2 = 0{,}9871$ (et R^2 ajusté $= 0{,}9855$) montre que près de 98,7 % de la variance totale de la consommation d'électricité est expliquée par la variation de l'année. Cela traduit une forte capacité explicative du modèle linéaire utilisé.

D'après les résultats ci-dessus, on a $F^* = 613{,}2 > F_{(1,8),0.05} = 5{,}32$ pour un seuil de signification classique $\alpha = 0{,}05$. Donc, on rejette l'hypothèse nulle H_0, ce qui confirme que le modèle est globalement significatif.

3.1.6 Prévision de l'année 2024/2025

Nous réalisons une prévision de la consommation d'électricité pour les années 2024 et 2025, dans le but d'anticiper les besoins énergétiques à venir et d'appuyer les prises de décision en matière de planification.

```
Prévisions pour 2024 et 2025 :
> for (i in 1:2) {
+   annee_pred <- nouvelles_annees$Annee[i]
+   cat("\nAnnée :", annee_pred, "\n")
+   cat("  Prévision :", round(pred_confiance[i,"fit"]), "\n")
+   cat("  IC 95%     : [", round(pred_confiance[i,"lwr"]), ",", round(pred_con$
+   cat("  IP 95%     : [", round(pred_prediction[i,"lwr"]), ",", round(pred_pr$
+ }

Année : 2024
  Prévision : 10034101
  IC 95%     : [ 9771790 , 10296412 ]
  IP 95%     : [ 9569074 , 10499129 ]

Année : 2025
  Prévision : 10488079
  IC 95%     : [ 10187657 , 10788501 ]
  IP 95%     : [ 10000537 , 10975620 ]
```

FIGURE 3.9 – Prévision de la consommation d'électricité : 2024–2025.

Les prévisions indiquent une augmentation de la consommation d'électricité en 2024 et 2025. Les intervalles de confiance à 95 % suggèrent une estimation précise de la moyenne attendue, tandis que les intervalles de prédiction traduisent l'incertitude associée aux valeurs futures individuelles.

Cette évolution est représentée sur le graphique suivant, qui illustre à la fois les valeurs observées, les projections futures et les marges d'erreur associées.

Pour visualiser les prévisions obtenues, nous avons ajouté les nouvelles années de prévision en créant un nouveau jeu de données : `nv_donnees <- data.frame(Annee = c(2024, 2025))`, puis nous avons tracé la série observée à l'aide de la commande :

`plot(nv_donnees$Annee, nv_donnees$Consommation_GWh)`.

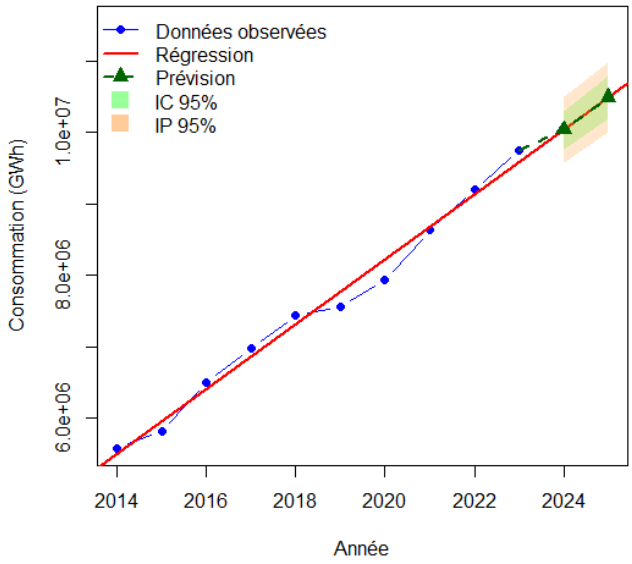


FIGURE 3.10 – Consomation d'électricité 2024/2025.

Les prévisions mettant en évidence une hausse de la consommation d'électricité en 2024 et 2025, il est essentiel d'anticiper cette tendance par des mesures appropriées. Parmi les décisions possibles : planifier l'extension des capacités de production, optimiser la gestion de la charge énergétique en période de pointe, investir dans les sources d'énergie renouvelable, et renforcer les dispositifs de maîtrise de la demande. Ces actions permettront d'assurer un équilibre durable entre l'offre et la demande énergétique à moyen terme.

3.2 Étude de la série chronologique par RLM

3.2.1 Présentation des données

Les données utilisées dans cette étude correspondent à la production mensuelle de blé, exprimée en milliers de tonnes (MT), en Algerie, sur la période de 2012 à 2021. Ces données ont été extraites de la base de (**ONS**)*Office National des Statistiques* .`https://www.ons.dz/`.

TABLE 3.2 – Production mensuelle de blé en Algérie.

Année	Jan	Fév	Mar	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc
2012	120	125	130	150	180	220	250	270	260	230	200	150
2013	125	130	135	155	185	225	255	275	265	235	205	155
2014	130	135	140	160	190	230	260	280	270	240	210	160
2015	135	140	145	165	195	235	265	285	275	245	215	165
2016	140	145	150	170	200	240	270	290	280	250	220	170
2017	145	150	155	175	205	245	275	295	285	255	225	175
2018	150	155	160	180	210	250	280	300	290	260	230	180
2019	155	160	165	185	215	255	285	305	295	265	235	185
2020	160	165	170	190	220	260	290	310	300	270	240	190
2021	165	170	175	195	225	265	295	315	305	275	245	195

3.2.2 Visualisation des données

Pour commencer l'analyse, nous visualisons la série temporelle à l'aide de la commande `plot(data$serie_blé)`, ce qui permet d'observer l'évolution mensuelle du production de blé.

Cette première représentation met en évidence une tendance croissante ainsi qu'un comportement saisonnier marqué dans la figure 3.11.

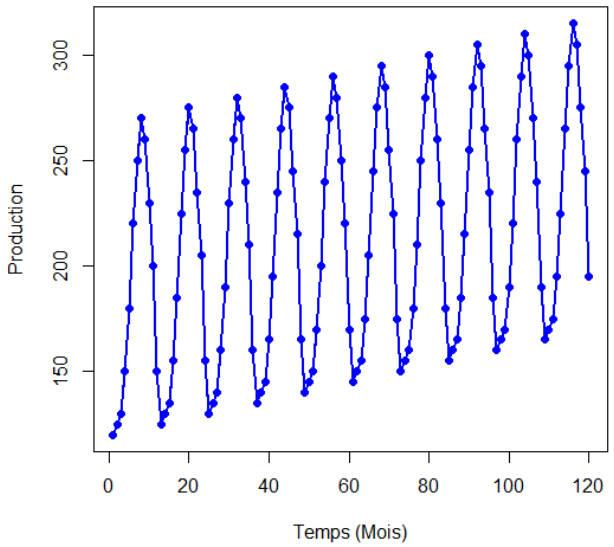


FIGURE 3.11 – Production mensuelle de blé (2012-2021).

Pour mieux comprendre la structure de la série, nous effectuons une décomposition en ses composantes fondamentales (tendance, saisonnalité et résidu) à l'aide de la commande decompose(). La commande `plot(decompose(serie_ts))` permet alors de visualiser séparément chacune de ces composantes, facilitant ainsi l'interprétation des dynamiques sous-jacentes à la série.

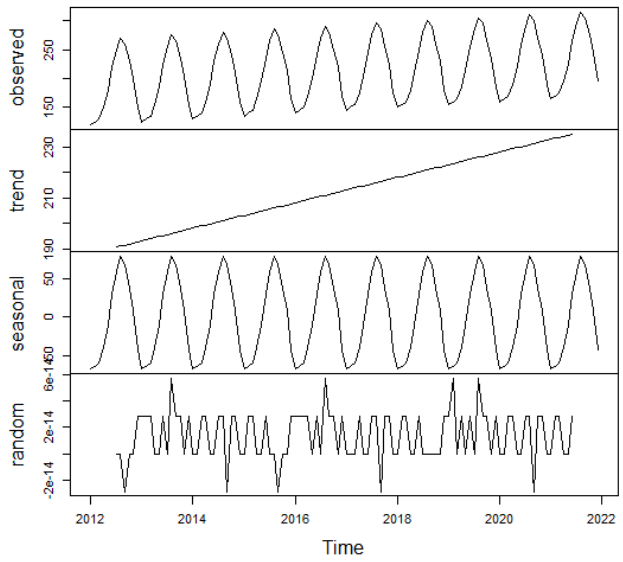


FIGURE 3.12 – Décomposition de la série.

La série met en évidence une tendance croissante marquée, accompagnée d'une saisonnalité régulière, avec des pics récurrents observés durant tout au long de l'année. Les résidus, quant à eux, sont faibles et ne présentent pas de structure particulière, ce qui confirme que la tendance et la saisonnalité capturent efficacement la dynamique de la série.

3.2.3 Ajustement du modèle de régression linéaire

Pour modéliser la production de blé, nous ajustons un modèle de régression linéaire prenant en compte à la fois la tendance temporel et la saisonnalité mensuelle. Le modèle s'écrit sous la forme :

$$Y_t = \beta_0 + \beta_1 t + \sum_{j=2}^{12} \lambda_j S_t^j + \varepsilon_t,$$

où

— Y_t représente la production de blé au temps t,

— t est la variable de temps (en mois),

— S_t^j sont des variables indicatrices (ou "dummies") pour les mois de l'année (le 12$^{\text{e}}$ mois servant de référence),

— ε_t est un terme d'erreur aléatoire.

L'estimation du modèle peut être réalisée dans le logiciel `R` avec la commande suivante :

`lm(serie_blé ~ Temps + Mois, data=data)`.

Ce modèle permet de quantifier l'effet de la tendance et des composantes saisonnières sur l'évolution de la production de blé, et d'envisager des prévisions plus fiables.

```
> modele

Call:
lm(formula = Serie_blé ~ Temps + Mois, data = data)

Coefficients:
(Intercept)        Temps        Mois2        Mois3        Mois4        Mois5
   119.5833       0.4167       4.5833       9.1667      28.7500      58.3333
      Mois6        Mois7        Mois8        Mois9       Mois10       Mois11
    97.9167     127.5000     147.0833     136.6667     106.2500      75.8333
     Mois12
    25.4167
```

FIGURE 3.13 – Ajustement du modèle sur les données observées.

Voici l'expression explicite du modèle de régression linéaire estimé à partir des résultats affichés :

$$\text{blé}_t = 119.5833 + 0.4167\ \text{Temps} + 4.5833S_t^2 + 9.1667S_t^3 + 28.7500S_t^4 + 58.3333S_t^5 + 97.9167S_t^6 + 127.5000S_t^7 + 147.0833S_t^8 + 136.6667S_t^9 + 106.2500S_t^{10} + 75.8333S_t^{11} + 25.4167S_t^{12}$$

Remarque 3.2.1

— *Le mois de janvier est utilisé comme modalité de référence, c'est pourquoi il n'apparaît pas explicitement dans le modèle. Les coefficients des autres mois représentent donc l'écart moyen par rapport à janvier.*

— *La variable* `Temps` *permet de capturer la tendance linéaire de la série, indiquant une augmentation régulière du production de blé au fil du temps.*

— *Les variables* `Mois` *sont des variables indicatrices (ou dummy variables) valant 1 lorsque l'observation concerne le mois k, et 0 sinon.*

— *Le terme ε_t désigne l'erreur aléatoire à l'instant t, supposée suivre une distribution normale centrée réduite si les hypothèses du modèle linéaire sont respectées.*

Pour visualiser la droite de régression ajustée au modèle estimé, on utilise la commande `abline(modele)`.

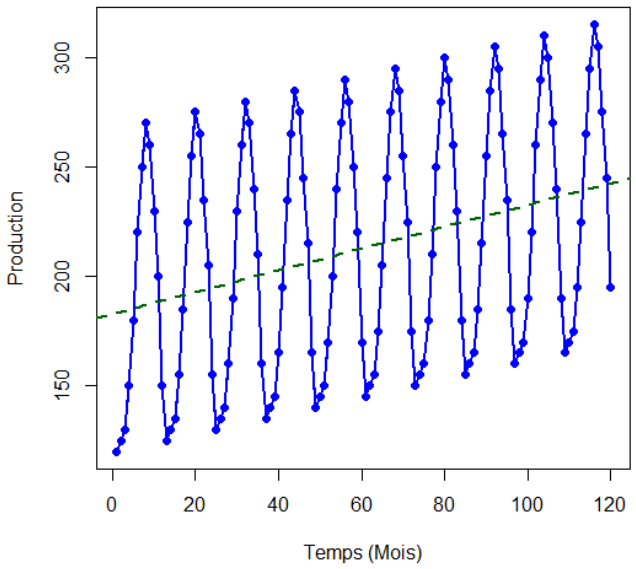


FIGURE 3.14 – La droite de l'équation de la série.

3.2.4 Évaluation du modèle

Test d'indépendance des résidus (ACF) :

Pour vérifier l'indépendance des erreurs, nous traçons la fonction d'autocorrélation (ACF) des résidus à l'aide de la commande `acf(residuals(modele))`.

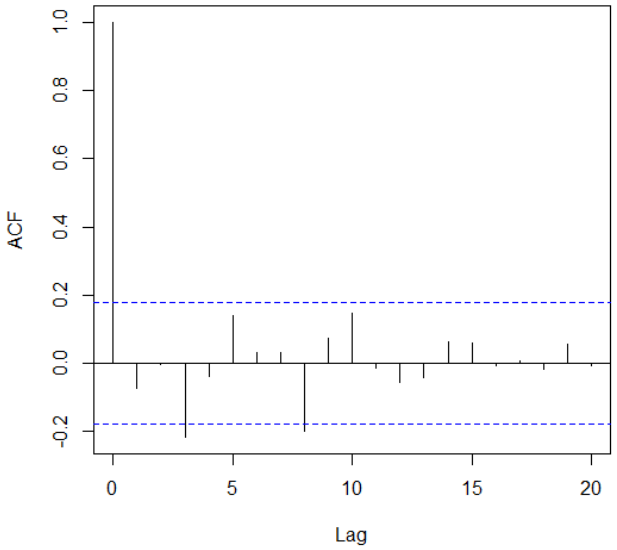


FIGURE 3.15 – ACF des résidus.

L'analyse de la fonction d'autocorrélation (ACF) des résidus montre que, hormis le décalage zéro, l'ensemble des autocorrélations se situent à l'intérieur des bandes de confiance. Cela indique une absence significative d'autocorrélation dans les résidus, ce qui suggère que le modèle ajusté capture correctement la structure temporelle des données et que l'hypothèse d'indépendance des erreurs est globalement vérifiée.

Test des résidus par rapport aux prédicteur :

Afin de détecter d'éventuelles tendances ou structures non prises en compte par le modèle, nous représentons graphiquement les résidus en fonction de la variable explicative `Temps` à l'aide de la commande `plot(data$Temps,residuals(modele))`.

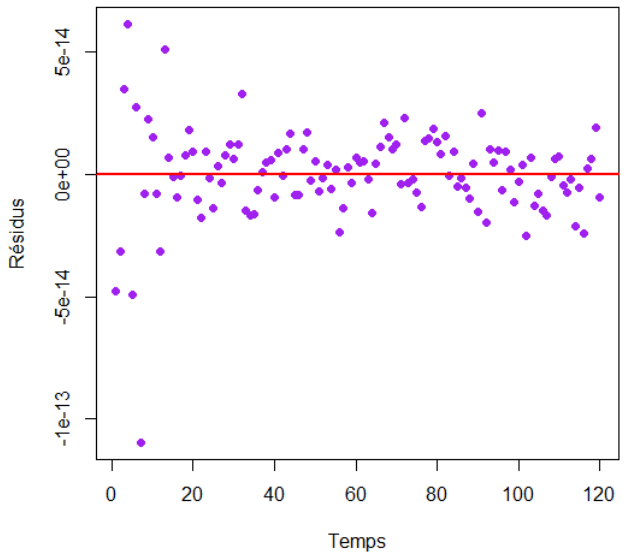


FIGURE 3.16 – Résidus en fonction de l'année et du mois.

L'analyse du nuage des points montre que les résidus sont globalement centrés autour de zéro et ne présentent pas de structure particulière ou de tendance apparente. Cette dispersion aléatoire des résidus suggère que le modèle ajusté ne laisse pas de composante temporelle non expliquée et que l'hypothèse d'indépendance des erreurs est respectée.

Test des résidus par rapport aux valeur ajustée

Nous utilisons la commande `plot(fitted(modele),residuals(modele))` pour examiner la relation entre les résidus et les valeurs ajustées.

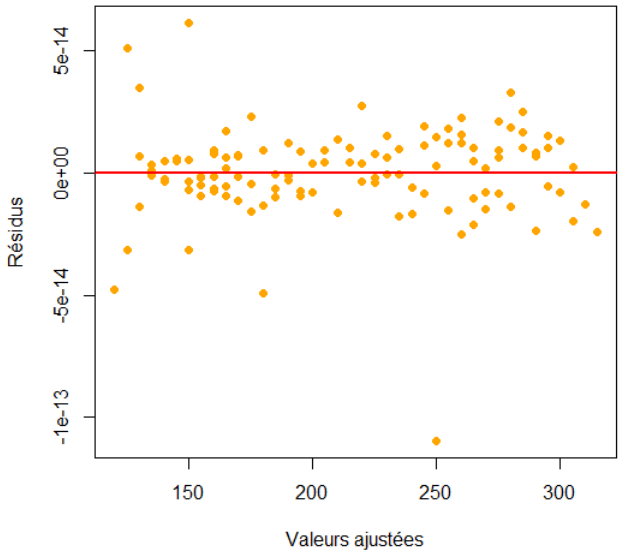


FIGURE 3.17 – Résidus en fonction de valeur ajustée.

On remarque que le nuage de points des résidus montre une dispersion homogène et aléatoire autour de la ligne zéro, sans motif systématique apparent. Cette répartition confirme l'hypothèse d'homoscédasticité et indique que la variance des erreurs reste constante quel que soit le niveau des prédictions.

Test de normalité des résidus :

La normalité des résidus est vérifiée à l'aide du test de Shapiro-Wilk, en utilisant la commande : `shapiro.test(residuals(modele))`.

```
> shapiro.test(residuals(modele))

        Shapiro-Wilk normality test

data:  residuals(modele)
W = 0.87449, p-value = 1.192e-08
```

FIGURE 3.18 – Test de normalité des résidus.

Le test de normalité de Shapiro-Wilk appliqué aux résidus donne une statistique $W = 0.874$ avec une p-valeur très faible ($p < 0.0001$). Cette p-valeur largement inférieure au seuil habituel de 5% conduit à rejeter l'hypothèse de normalité des résidus. Ainsi, les résidus ne suivent pas une distribution normale, ce qui constitue une violation de l'une des hypothèses classiques du modèle de régression linéaire.

Cependant, cette violation ne remet pas nécessairement en cause la validité globale du modèle pour la prévision. En effet, tant que les autres hypothèses des résidus sont raisonnablement respectées, le modèle reste exploitable pour produire des prévisions.

3.2.5 Validation du modèle

La validation du modèle repose sur le test global de significativité basé sur la statistique de Fisher (F-statistic). Cette statistique permet de tester l'hypothèse nulle H_0 selon laquelle aucun des prédicteurs n'a d'effet significatif sur la variable dépendante.

On compare la valeur observée de la statistique F^* à la valeur critique F_{tab} de la loi de Fisher, aux degrés de liberté $k - 1$ (numérateur) et $L - m - p$ (dénominateur). Si $F^* > F_{\text{tab}}$, on rejette l'hypothèse nulle, ce qui indique que le modèle est globalement significatif.

où k est le nombre total de paramètres estimés, soit : $k =$ tendance + saisonnalité.

Cette validation est effectuée à l'aide de la commande `summary(modele)` dans R, qui fournit notamment la valeur de F, le degré de signification, ainsi que les coefficients estimés et leurs tests associés.

```
Residuals:
       Min         1Q     Median         3Q        Max
-1.096e-13 -7.960e-15  9.600e-17  9.480e-15  6.148e-14

Coefficients:
             Estimate Std. Error   t value Pr(>|t|)
(Intercept) 1.196e+02  6.922e-15 1.728e+16   <2e-16 ***
t           4.167e-01  5.266e-17 7.913e+15   <2e-16 ***
Mois2       4.583e+00  8.892e-15 5.155e+14   <2e-16 ***
Mois3       9.167e+00  8.892e-15 1.031e+15   <2e-16 ***
Mois4       2.875e+01  8.893e-15 3.233e+15   <2e-16 ***
Mois5       5.833e+01  8.894e-15 6.559e+15   <2e-16 ***
Mois6       9.792e+01  8.895e-15 1.101e+16   <2e-16 ***
Mois7       1.275e+02  8.897e-15 1.433e+16   <2e-16 ***
Mois8       1.471e+02  8.899e-15 1.653e+16   <2e-16 ***
Mois9       1.367e+02  8.902e-15 1.535e+16   <2e-16 ***
Mois10      1.062e+02  8.904e-15 1.193e+16   <2e-16 ***
Mois11      7.583e+01  8.907e-15 8.514e+15   <2e-16 ***
Mois12      2.542e+01  8.910e-15 2.852e+15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.988e-14 on 107 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 7.589e+31 on 12 and 107 DF,  p-value: < 2.2e-16
```

FIGURE 3.19 – Résumé statistique du modèle de régression linéaire multiple.

L'ajustement du modèle est excellent. Les résidus sont quasi nuls, indiquant une très bonne concordance entre les valeurs observées et ajustées. Tous les coefficients sont hautement significatifs ($p < 2 \times 10^{-16}$), confirmant leur contribution importante. Le coefficient de détermination $R^2 = 1$ montre que le modèle explique parfaitement la variabilité des données. La statistique de Fisher très élevée ($F = 7.589 \times 10^{31}$) confirme la significativité globale du modèle. L'erreur standard des résidus est très faible (1.988×10^{-14}), ce qui témoigne de la précision de l'ajustement.

Ces résultats confirment la bonne qualité d'ajustement du modèle et sa pertinence pour la prévision du production de blé mensuel.

3.2.6 Prévision de l'année 2022/2023

Nous réalisons une prévision de la production de blé en Algérie pour les années 2022 et 2023. L'objectif est d'anticiper l'évolution probable de la production dans les années à venir, afin de mieux planifier les ressources, de fournir des éléments d'aide à la décision aux acteurs économiques, agricoles et politiques, et d'anticiper d'éventuelles fluctuations du marché.

Les valeurs prédites ont été obtenues en appliquant la commande `predict()` aux nouvelles observations correspondant aux différentes combinaisons de temps et de mois. Le tableau suivant présente les prévisions pour les 24 mois à venir.

```
> predictions
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
170 175 180 200 230 270 300 320 310 280 250 200 175 180 185 205 235 275 305 325
 21  22  23  24
315 285 255 205
```

FIGURE 3.20 – Prévision de la production de blé : 2022-2023.

Ces prévisions mettent en évidence la poursuite de la tendance haussière observée au cours des années précédentes. Cette information est essentielle pour les décideurs, car elle leur permet d'adapter leurs stratégies en fonction des besoins anticipés en matière de production de blé.

Pour visualiser les prévisions obtenues, nous avons créé un nouveau jeu de données :

```
nv_donnees <- data.frame(Annee = nouvelle_annee, Mois = factor(nouveau_mois,
levels = levels(data$Mois)), Temps = nouveau_Temps),
```

puis nous avons tracé la série observée à l'aide de la commande :

```
plot(nv_donnees$Temps, predictions).
```

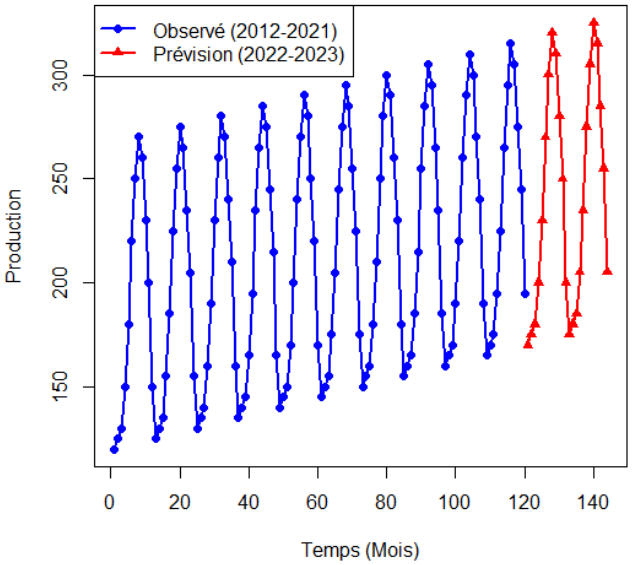


FIGURE 3.21 – Production de blé 2022/2023.

À partir de ce graphe de prévision, on observe une tendance générale haussière accompagnée d'une forte saisonnalité régulière continue de production de blé pour les deux années à venir. Les prévisions indiquent la poursuite de cette dynamique avec des volumes de production en augmentation. Ces résultats sont particulièrement importants pour les décideurs, car ils soulignent la nécessité d'adapter les infrastructures et les politiques agricoles. D'une part, il devient essentiel d'étendre les capacités de stockage par l'agrandissement des silos et entrepôts pour absorber les récoltes supplémentaires. D'autre part, l'adaptation des infrastructures de transformation notamment les usines d'amidon, de biocarburants et d'alimentation animale permettra de valoriser cette matière première excédentaire. En parallèle, l'organisation logistique des transports devra être optimisée, en mobilisant les réseaux routier, ferroviaire et maritime afin de fluidifier l'écoulement des volumes accrus vers les marchés nationaux et internationaux. Par ailleurs, une gestion efficace des ressources humaines agricoles, avec une planification adaptée de la main-d'œuvre qualifiée aux différentes phases de production, de récolte et de transformation, s'avérera indispensable. Enfin, la sécurisation des marchés et la diversification des débouchés par le développement d'accords commerciaux contribueront à garantir l'écoulement des surplus de production. Ainsi, ces prévisions constituent un outil stratégique essentiel pour l'élaboration de politiques agricoles et économiques cohérentes avec les perspectives de croissance du secteur.

Conclusion

Ce mémoire a exploré l'utilisation de la régression linéaire dans le cadre de la prévision des séries chronologiques, en mettant en lumière son efficacité ainsi que ses limites face à la complexité des données temporelles.

Dans un premier temps, nous avons présenté les fondements théoriques des séries chronologiques et les mécanismes qui les sous-tendent, tels que la stationnarité, la tendance, la saisonnalité et l'autocorrélation. Ces concepts ont permis de mieux comprendre la structure des phénomènes étudiés et d'orienter le choix des modèles de prévision.

La deuxième partie du travail a été consacrée à l'étude des modèles de régression linéaire simples, multiples et temporels, avec une attention particulière portée aux hypothèses sous-jacentes, aux méthodes d'évaluation et aux diagnostics de validation. L'analyse des composantes (tendance et saisonnalité), ainsi que les prévisions qui en découlent, ont montré que la régression linéaire, bien qu'intuitive et facile à mettre en œuvre, peut offrir des résultats satisfaisants lorsque les données présentent une structure suffisamment régulière.

Enfin, la partie pratique a permis d'appliquer concrètement les modèles étudiés à des séries réelles. Les résultats obtenus mettent en évidence une tendance haussière significative, ainsi qu'une saisonnalité annuelle marquée. Les modèles ajustés ont démontré une bonne capacité explicative, confirmant la pertinence de la régression dans un cadre prédictif. Toutefois, certaines limites ont été identifiées, notamment une hétéroscédasticité dans certains cas, ce qui remet en question les hypothèses classiques du modèle linéaire.

Malgré ces limites, la régression linéaire constitue une approche intéressante pour débuter l'analyse de séries chronologiques, notamment en contexte exploratoire. Néanmoins, pour des phénomènes plus complexes ou pour améliorer la précision des prévisions, il serait pertinent d'envisager des modèles plus avancés, tels que les modèles ARIMA ou SARIMA, les méthodes basées sur les réseaux de neurones récurrents (RNN) ou les modèles hybrides combinant approche linéaire et non linéaire. Ces méthodes permettent de mieux capturer les dynamiques sous-jacentes aux séries, en tenant compte des comportements non linéaires, des effets de mémoire longue ou encore des changements structurels dans les données.

Bibliographie

[1] Abd Elfattah, M. A., Elshafei, A. L., & Morsy, M. M. (2018). Forecasting electricity consumption using linear regression and neural networks. *International Journal of Engineering Research and Applications*, 8(3), 47–52.

[2] Akrour, Y. (2022). *Prévision de la demande d'électricité par régression linéaire et réseaux de neurones artificiels*. Mémoire de maîtrise, Université du Québec à Trois-Rivières.

[3] Arnaud, R. (2022). *Séries chronologiques*. Université de Bourgogne.

[4] Brockwell, P. J., & Davis, R. A. (2006). *Introduction to Time Series and Forecasting* (2e éd.). Springer.

[5] Charpentier, A. (2022). *Cours de séries temporelles – Théorie et applications*. Université Paris-Dauphine.

[6] Coutrot, B., & Droesbeke, J. J. (1990). *Les Méthodes de prévision*. Presses Universitaires de France.

[7] Cowpertwait, P. S. P., & Metcalfe, A. V. (2009). *Introductory Time Series with R*. Springer.

[8] Girard, Y. (2011). *Séries chronologiques : Méthodes classiques et modèles à base de copules*. Mémoire de maîtrise, Université du Québec à Trois-Rivières.

[9] Haurie, A. (1966). Recherche opérationnelle : Les modèles linéaires et les séries chronologiques. *L'Actualité économique*, HEC Montréal.

[10] Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting : Principles and Practice* (3e éd.). OTexts, Monash University. `https://otexts.com/fpp3/`

[11] Irawan, D., & Abdillah, L. A. (2016). Forecasting time series data using multiple linear regression model. *Journal of Theoretical and Applied Information Technology*, 89(1), 123–130.

[12] Mokhtari, F. *Séries chronologiques : Cours, exercices et travaux pratiques*. Université Dr Moulay Tahar, Saïda.

[13] Moore, D. S. (2021). *Chapitre 4 – Corrélation et régression linéaire simple*. Université de Pittsburgh.

[14] Shumway, R. H., & Stoffer, D. S. (2017). *Time Series Analysis and Its Applications : With R Examples* (4e éd.). Springer. `https://www.stat.pitt.edu/stoffer/tsa4/`

[15] Stafford, J., & Sarrasin, B. (2005). *La prévision-prospective en gestion : Tourisme, Loisir, Culture*. Presses de l'Université du Québec.

[16] Suryanarayana, M. A., & Naik, S. (2020). Time series forecasting using linear regression : a case study on COVID-19 data. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 6(3), 35–41.

[17] Thibodeau, K. (2011). *Application de la méthode Box-Jenkins aux séries chronologiques*. Mémoire de maîtrise, Université du Québec à Trois-Rivières.

Annexe A

A. Détails techniques des tests

A.1 Test de signification globale (ANOVA)

On peut tester si la régression multiple à k variables indépendantes est significative dans son ensemble.

En se basant sur la relation SCT = SCE + SCR de décomposition de la variance, on peut établir un test permettant de vérifier la significativité globale du modèle.

Lorsque toutes les hypothèses du modèle sont satisfaites et sous l'hypothèse H_0 :

$$\begin{cases} H_0 : \text{tous les } \beta_j = 0, \\ H_1 : \exists\, \beta_j \neq 0 \quad \text{hypothèse de significativité globale.} \end{cases}$$

Pour ce faire, nous utiliserons le rapport F^* du carré moyen expliqué sur le carré moyen résiduel. Ce rapport devrait suivre une loi de Fisher sous H_0 (rapport de deux chi-deux).

$$F^* = \frac{CME}{CMR} = \frac{\text{SCE}/(k)}{\text{SCR}/(L-k-1)} \sim \mathcal{F}(k, L-k-1).$$

Notons que la p-valeur associée à cette statistique est :

$$\text{p-value} = P(F_{\text{tab}} > F^* \mid H_0 \text{ vraie}).$$

L'hypothèse de normalité des erreurs ($\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$) implique que sous l'hypothèse H_0, F^* suit une loi de Fisher avec $(k, L-k-1)$ degrés de liberté.

Pour un seuil critique α (généralement 5%), nous comparons ce F^* calculé avec le F théorique (de la table). Alors le test se résume si $F^* > F_{\text{tab}}$, nous rejetons l'hypothèse H_0 et donc le modèle est globalement explicatif (significatif) dans son ensemble. Sinon, est considérée comme non significative.

A.2 Tests de contribution marginale (test t de student)

On peut s'intéresser à savoir si la contribution de chaque variable explicative est significative. Il s'agit de tester :

$$H_0 : \beta_j = 0, \quad \text{contre} \quad H_1 : \beta_j \neq 0, \quad \text{pour} \quad j = 1, \ldots, k.$$

À cette fin, nous utilisons la statistique de test suivante :

$$t^* = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim \mathcal{T}(L-k-1).$$

Ce rapport suit une loi de Student à $L-k-1$ degrés de liberté.

Pour calculer $\hat{\sigma}_{\hat{\beta}_j}$, on s'en remet à un logiciel de statistique mais retenons que :

$$\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}, \quad \text{d'où} \quad \hat{\sigma}^2_{\hat{\beta}} = \hat{\sigma}^2 (X'X)^{-1}.$$

alors que :

$$\hat{\sigma}^2 = \frac{1}{L-k-1}(Y'Y - \hat{\beta}'X'Y).$$

On rejette H_0 si $|t^*| > St_{L-k-1,\alpha/2}$ (de la table) et on conclut que la variable explicative testée est significative au seuil α.

Notons que la p-valeur associée à cette statistique est :

$$\text{p-value} = P(t_{tab} > t^* \mid H_0 \text{ vraie}).$$

Annexe B

B. Homoscédasticité et Hétéroscédasticité

Dans un modèle de régression linéaire, une des hypothèses fondamentales porte sur la variance des erreurs. On suppose généralement que les erreurs sont **homoscédastiques**, c'est-à-dire qu'elles ont une variance constante à travers toutes les observations. Cette propriété est essentielle pour assurer la validité des estimateurs et des tests statistiques classiques (tests t, F, etc.).

B.1 Homoscédasticité

La homoscédasticité désigne une situation où les erreurs aléatoires du modèle de régression ont une **variance constante**, quelle que soit la valeur de la variable indépendante. Autrement dit, les points sont répartis de manière régulière autour de la droite de régression. Cela garantit une estimation efficace des coefficients du modèle.

Représentation graphique : dans un nuage de points des résidus, l'homoscédasticité se manifeste par une dispersion constante.

B.2 Hétéroscédasticité

L'hétéroscédasticité, en revanche, apparaît lorsque la variance des erreurs **n'est pas constante**. Elle peut augmenter ou diminuer selon la valeur des variables indépendantes. Ce phénomène viole les hypothèses classiques de la régression linéaire (notamment dans les moindres carrés ordinaires).

Conséquences :

— Les estimateurs des coefficients restent non biaisés, mais ils deviennent inefficaces.

— Les tests statistiques (valeurs p, intervalles de confiance) peuvent être faussés.

Représentation graphique : l'hétéroscédasticité se traduit souvent par une forme en éventail ou en cône dans le graphe des résidus.

B.3 Détection

Quelques méthodes courantes pour détecter l'hétéroscédasticité :

— Visualisation graphique : résidus vs. valeurs ajustées.

— Test de Breusch–Pagan ou test de White.

— Statistiques de résidus studentisés.

B.4 Traitement

Pour corriger l'hétéroscédasticité, plusieurs approches sont possibles :

— Transformation des variables (ex. : logarithmes).

— Modèles robustes à l'hétéroscédasticité.

— Régression pondérée (WLS).

— Utilisation de modèles adaptés pour séries chronologiques avec variances conditionnelles (ex. modèles ARCH/GARCH).

Annexe C

C. Tables statistiques

C.1 Table de la loi de Fisher-Snedecor

TABLE 3 – Table de la loi de Fisher pour $\alpha = 0{,}05$ (5 %).

$v_2 \backslash v_1$	1	2	3	4	5	6	8	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	238.9	243.9	249.0	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.37	19.41	19.45	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.84	8.74	8.65	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.04	5.91	5.78	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.82	4.67	4.53	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.14	3.98	3.82	3.63
7	5.59	4.74	4.35	4.12	3.97	3.87	3.73	3.57	3.41	3.23
8	5.32	4.46	4.07	3.84	3.68	3.58	3.44	3.27	3.10	2.92
9	5.12	4.26	3.87	3.63	3.46	3.37	3.23	3.05	2.87	2.69
10	4.96	4.10	3.71	3.48	3.31	3.22	3.07	2.89	2.71	2.53
11	4.84	3.98	3.59	3.36	3.18	3.09	2.94	2.76	2.59	2.41
12	4.75	3.88	3.49	3.26	3.11	3.00	2.85	2.69	2.52	2.34
13	4.67	3.81	3.41	3.18	3.02	2.91	2.76	2.60	2.42	2.24
14	4.60	3.74	3.34	3.11	2.95	2.84	2.69	2.52	2.34	2.16
15	4.54	3.68	3.29	3.06	2.90	2.79	2.64	2.46	2.29	2.11
16	4.49	3.63	3.24	3.01	2.85	2.74	2.59	2.41	2.23	2.06
17	4.45	3.59	3.20	2.98	2.81	2.70	2.55	2.37	2.19	2.02
18	4.41	3.55	3.16	2.94	2.77	2.66	2.51	2.33	2.15	1.98
19	4.38	3.52	3.13	2.90	2.74	2.62	2.47	2.29	2.11	1.95
20	4.35	3.49	3.10	2.87	2.71	2.60	2.45	2.26	2.08	1.92
21	4.32	3.47	3.07	2.84	2.68	2.56	2.42	2.23	2.06	1.90
22	4.30	3.44	3.05	2.82	2.66	2.54	2.40	2.21	2.04	1.88
23	4.28	3.42	3.03	2.80	2.64	2.52	2.38	2.18	2.01	1.86
24	4.26	3.40	3.01	2.78	2.62	2.50	2.36	2.16	1.99	1.84
25	4.24	3.38	2.99	2.76	2.60	2.48	2.34	2.14	1.97	1.83
30	4.17	3.33	2.94	2.69	2.54	2.42	2.27	2.07	1.90	1.76
40	3.98	3.22	2.84	2.59	2.43	2.31	2.16	1.96	1.79	1.64
60	3.78	3.10	2.74	2.47	2.31	2.18	2.01	1.83	1.66	1.51
120	3.92	3.07	2.68	2.45	2.29	2.17	2.01	1.83	1.65	1.50
∞	3.84	2.99	2.60	2.37	2.21	2.09	1.94	1.75	1.52	1.00

C.2 Table de la loi de Student

TABLE 4 – Table de la loi de student.

$n\backslash\alpha$	90%	80%	70%	60%	50%	40%	30%	20%	10%	5%	2%	1%
1	0.1584	0.3249	0.5095	0.7265	1.0000	1.3764	1.9626	3.0777	6.3138	12.7062	31.8205	63.6567
2	0.1421	0.2887	0.4447	0.6172	0.8165	1.0607	1.3862	1.8856	2.9200	4.3027	6.9646	9.9248
3	0.1366	0.2767	0.4242	0.5844	0.7649	0.9785	1.2498	1.6377	2.3534	3.1824	4.5407	5.8409
4	0.1338	0.2707	0.4142	0.5686	0.7407	0.9410	1.1896	1.5332	2.1318	2.7764	3.7469	4.6041
5	0.1322	0.2672	0.4082	0.5594	0.7267	0.9195	1.1558	1.4759	2.0150	2.5706	3.3649	4.0321
6	0.1311	0.2648	0.4043	0.5534	0.7176	0.9057	1.1342	1.4398	1.9432	2.4469	3.1427	3.7074
7	0.1303	0.2632	0.4015	0.5491	0.7111	0.8960	1.1192	1.4149	1.8946	2.3646	2.9980	3.4995
8	0.1297	0.2619	0.3995	0.5459	0.7064	0.8889	1.1081	1.3968	1.8595	2.3060	2.8965	3.3554
9	0.1293	0.2610	0.3979	0.5435	0.7027	0.8834	1.0997	1.3830	1.8331	2.2622	2.8214	3.2498
10	0.1289	0.2602	0.3966	0.5415	0.6998	0.8791	1.0931	1.3722	1.8125	2.2281	2.7638	3.1693
11	0.1286	0.2596	0.3956	0.5399	0.6974	0.8755	1.0877	1.3634	1.7959	2.2010	2.7181	3.1058
12	0.1283	0.2590	0.3947	0.5386	0.6955	0.8726	1.0832	1.3562	1.7823	2.1788	2.6810	3.0545
13	0.1281	0.2586	0.3940	0.5375	0.6938	0.8702	1.0795	1.3502	1.7709	2.1604	2.6503	3.0123
14	0.1280	0.2582	0.3933	0.5366	0.6924	0.8681	1.0763	1.3450	1.7613	2.1448	2.6245	2.9768
15	0.1278	0.2579	0.3928	0.5357	0.6912	0.8662	1.0753	1.3406	1.7531	2.1314	2.6025	2.9467
16	0.1277	0.2576	0.3923	0.5350	0.6901	0.8647	1.0711	1.3368	1.7459	2.1199	2.5835	2.9208
17	0.1276	0.2573	0.3919	0.5344	0.6892	0.8633	1.0690	1.3334	1.7396	2.1098	2.5669	2.8982
18	0.1274	0.2571	0.3915	0.5338	0.6884	0.8620	1.0672	1.3304	1.7341	2.1009	2.5524	2.8784
19	0.1274	0.2569	0.3912	0.5333	0.6876	0.8610	1.0655	1.3277	1.7291	2.0930	2.5395	2.8609
20	0.1273	0.2567	0.3909	0.5329	0.6870	0.8600	1.0640	1.3253	1.7247	2.0860	2.5280	2.8453
21	0.1272	0.2566	0.3906	0.5325	0.6864	0.8591	1.0627	1.3232	1.7207	2.0796	2.5176	2.8314
22	0.1271	0.2564	0.3904	0.5321	0.6858	0.8583	1.0614	1.3212	1.7171	2.0739	2.5083	2.8188
23	0.1271	0.2563	0.3902	0.5317	0.6853	0.8575	1.0603	1.3195	1.7139	2.0687	2.4999	2.8073
24	0.1270	0.2562	0.3900	0.5314	0.6848	0.8569	1.0593	1.3178	1.7109	2.0639	2.4922	2.7969
25	0.1269	0.2561	0.3898	0.5312	0.6844	0.8562	1.0584	1.3163	1.7081	2.0595	2.4851	2.7874
26	0.1269	0.2560	0.3896	0.5309	0.6840	0.8557	1.0575	1.3150	1.7056	2.0555	2.4786	2.7787
27	0.1268	0.2559	0.3894	0.5306	0.6837	0.8551	1.0567	1.3137	1.7033	2.0518	2.4727	2.7707
28	0.1268	0.2558	0.3893	0.5304	0.6834	0.8546	1.0560	1.3125	1.7011	2.0484	2.4671	2.7633
29	0.1268	0.2557	0.3892	0.5302	0.6830	0.8542	1.0553	1.3114	1.6991	2.0452	2.4620	2.7584