Analyse exploratoire des données - Devoir III

Quotients de mortalité,Tables de mortalité,ACP,ACC,SVD,Modèle de Lee-Carter.

Duc Duong Nguyen & Mohamed-Amine Bousahih

2020-05-26

Contents

Introduction

L'étude de la structure et de l'évolution d'une population dans une région est appelée démographie. La démographie est l'étude du nombre, de la répartition, du territoire et de la structure de la population et de ses changements, dans laquelle les changements se produisent en raison de la naissance (fertilité), du décès (mortalité) et de la migration.

L'objet principal de ce travail de recherche est les quotients de mortalité. Un quotient de mortalité, d'après une définition de l'INED, est une "probabilité, pour les personnes survivantes à un âge, de décéder avant l'âge suivant".

Le fil conducteur de ce travail de recherche sera une comparaison reposant sur ces quotients de mortalités entre les Etats-Unis et l'Espagne.

Nous observerons l'évolution de ces quotients de mortalité à travers le temps, en fonction du sexe et pour des groupes d'âges précis. Ceci nous permettra de distinguer des tendances mondiales et voir si en fonction de l'âge, les quotients de mortalité partagent une forme commune.

Comme tout bon statisticien, nous utiliserons une ACP (Analyse en composante principale) et une ACC (Analyse correspondance canonique) pour explorer des relations pouvant exister entre des groupes de variables que nous essayerons d'identifier au sein de ces données.

A l'issue de cette phase d'exploration, nous chercherons à effectuer des prédictions de ces quotients de mortalité. Pour effectuer ces prédictions, de nombreuses méthodes ont été proposées pour décrire le comportement des quotients de mortalité en fonction de l'âge, du temps pour une région donnée. En particulier, nous présenterons un modèle mathématique étant largement reconnu et utilisé proposé par Lee et Carter deux américains, qui en 1992, ont mis en place une méthode afin d'ajuster et de prévoir les taux de mortalité humaine.

Les données

L'ensemble de données qui sera utilisé est la base de données sur la mortalité humaine (HMD, Human Mortality Database organization), qui fournit un accès gratuit aux données historiques de mortalité pour des pays européens : France, Grande-Bretagne – enfaite l'Angleterre et le Pays de Galles – , Italie, Pays-Bas, Espagne, Suède. Couvrant également les Etats-Unis. Le HMD est une importante collection de données détaillées, cohérentes et de haute qualité sur la mortalité humaine.

Les tables de données sont téléchargés à partir de [https://www.mortality.org] (https://www.mortality.org).

Nous chargeons les tables de mortalité pour une année donné, pour les femmes, les hommes et l'ensemble de la population pour les différents pays. Nous obtenons une table universelle construite en fusionnant ces tables de mortalité.

```
## Observations: 252,780
## Variables: 12
## $ Year    <int> 1816, 1816, 1816, 1816, 1816, 1816, 1816, 1816, 1816, 1816,...
## $ Age     <int> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
## $ mx      <dbl> 0.18699, 0.04670, 0.03393, 0.02291, 0.01599, 0.01383, 0.012...
## $ qx      <dbl> 0.16573, 0.04564, 0.03336, 0.02265, 0.01587, 0.01374, 0.012...
## $ ax      <dbl> 0.31, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50, 0.50,...
## $ lx      <int> 100000, 83427, 79620, 76963, 75220, 74026, 73009, 72131, 71...
## $ dx      <int> 16573, 3807, 2656, 1743, 1194, 1017, 878, 748, 633, 535, 44...
## $ Lx      <int> 88633, 81523, 78291, 76092, 74623, 73518, 72570, 71757, 710...
## $ Tx      <int> 4111543, 4022910, 3941387, 3863095, 3787003, 3712380, 36388...
## $ ex      <dbl> 41.12, 48.22, 49.50, 50.19, 50.35, 50.15, 49.84, 49.44, 48....
## $ Gender  <fct> F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F, F,...
## $ Country <fct> France, France, France, France, France, France, France, Fra...
```

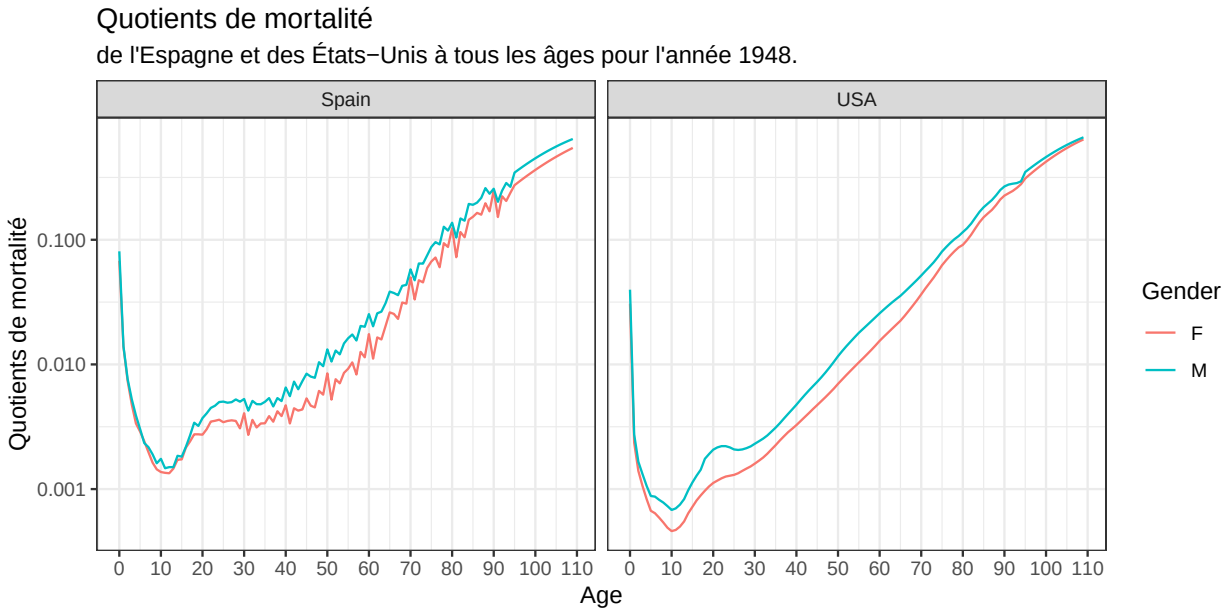
Pour mener à bien ce travail de recherche, nous utiliserons six colonnes de cette table universelle : Year, Age, mx , ex , Gender et Country.

- mx représente un quotient de mortalité pour un âge x et une année t. Il se calcule en divisant les décès à un âge X par les survivants à un âge X.

- ex représente l'espérance de vie.

Visualisations et réarrangement de nos données

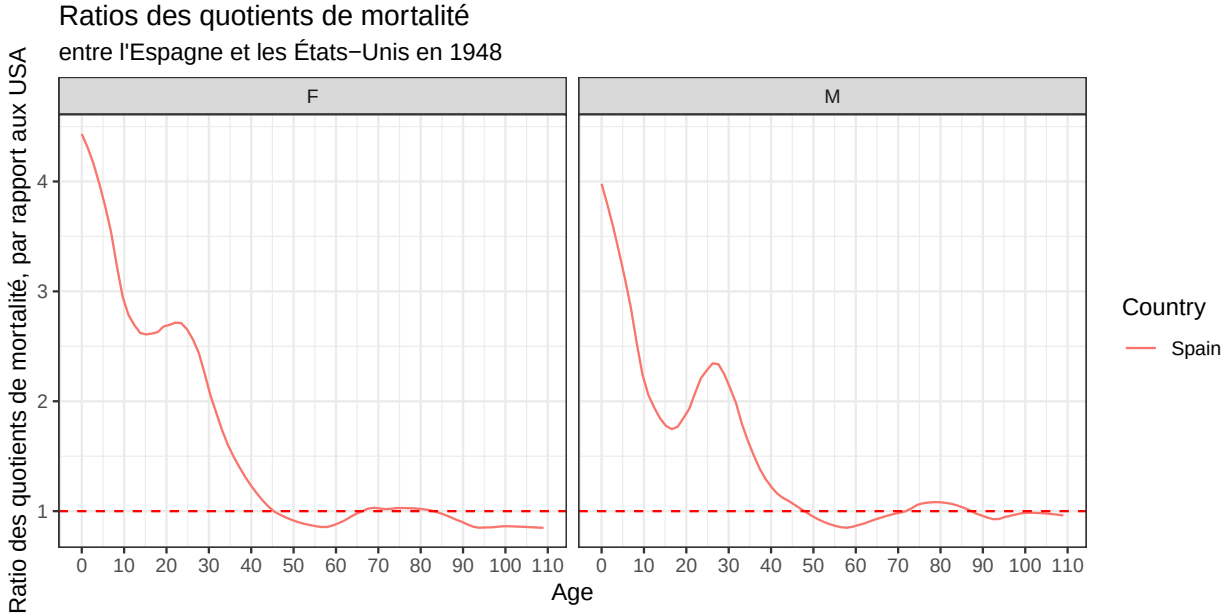
Visualisations des quotients de mortalité en fonction de l'âge, des années, des pays.



Ce graphique présente les taux centraux de mortalité par âge et par sexe pour l'Espagne et les États-Unis en 1948.

- On peut observer, à tous les âges, pour l'Espagne et les États-Unis, que le risque de mourir des hommes est plus élevé que celui des femmes. De plus, pour les âges entre 0 et 50 ans, le risque de mourir en Espagne est plus élevé que celui des États-Unis.
- La plupart de ces décès surviennent immédiatement après la naissance. Pour l'Espagne, le risque annuel de mourir à la naissance est de 68 enfants pour 10.000 pour les filles et de 80 enfants pour 10.000 pour les garçons.Pour les États-Unis, le risque annuel de mourir à la naissance est de 30 enfants pour 10.000 pour les filles et de 40 enfants pour 10.000 pour les garçons.
- De la naissance au premier anniversaire, la probabilité de mourir diminue fortement pour les deux sexes et les deux pays.
- Pour les deux sexes, à partir de 1 ans, la probabilité de mourir en Espagne et aux États-Unis diminue progressivement, atteignant un risque minimum à l'âge de 10 ans et/ou 11 ans.
- Le risque de mourir augmente fortement à l'adolescence.
- L'excès de mortalité masculine est le plus élevé entre 20 et 30 ans, période au cours de laquelle le risque de mourir est de $1,6$ plus élevé chez les hommes que chez les femmes du même âge.
- Peu importe le sexe, aux États-Unis, nous observons une croissance exponentielle qui continue de croître pour les aînés. Pour l'Espagne, les courbes sont en forme de "dent de scie" avec des pics de mortalité à des âges spécifiques : 65 ans, 70 ans, 80 ans dépassant la mortalité des USA.

- À partir de l'âge de 80+, nous pensons que l'estimation exacte est compliquée notamment dû à des erreurs de recensement et, pour les personnes extrêmement âgées (105 et plus), par le faible nombre de personnes vivantes.



- Nous constatons qu'un écart de mortalité se creuse entre les deux pays à la naissance. En effet, en Espagne, le risque annuel de décès à la naissance pour les filles est 4.43 supérieur à celui des États-Unis. Concernant les garçons, ce risque annuel de décès est supérieur à 3.97 à celui des États-Unis.

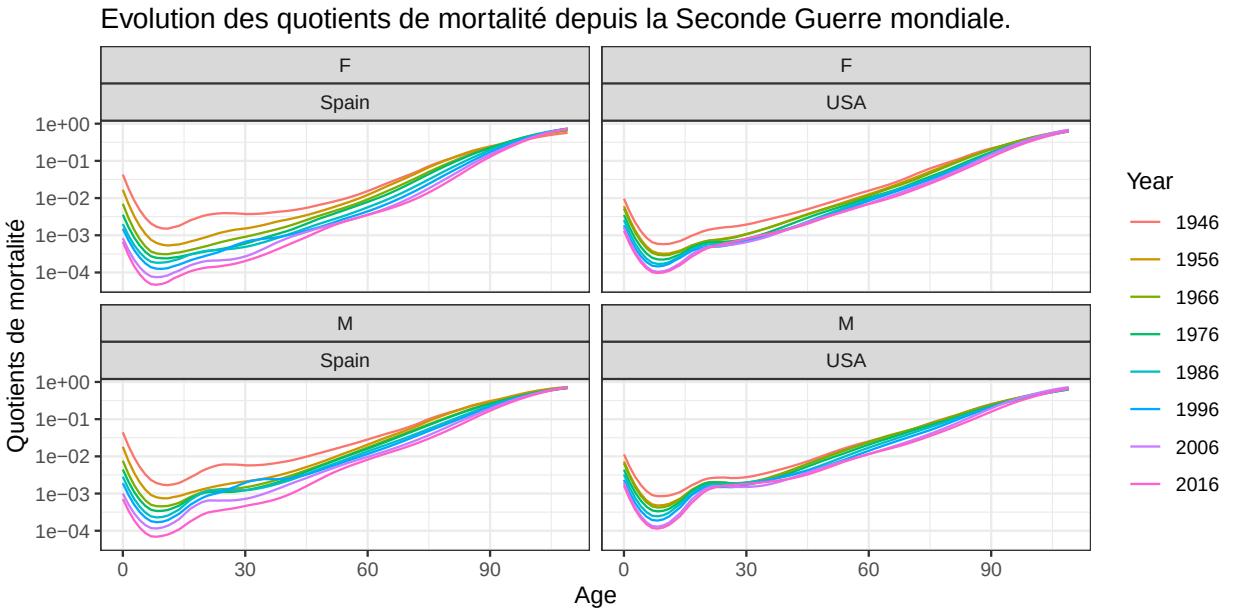
- De la naissance au 16ème anniversaire, ce ratio diminue fortement pour les deux sexes signifiant une baisse d'écart de mortalité entre les deux pays. Pour les garçons, ce ratio atteint un minimum local à 16 ans signifiant que le risque annuel de décès à cet âge pour les garçons est 1.74 supérieur à celui des États-Unis. Concernant les filles, le risque annuel de décès en ce point de minimum local est 2.61 supérieur à celui des États-Unis.

- Comme nous l'avons remarqué précédement, le risque de mourir augmente fortement à l'adolescence. Cela a pour conséquence que l'écart de mortalité entre les deux pays se creuse impliquant une augmentation des ratios des quotients de mortalité entre 16 ans et 23 ans.

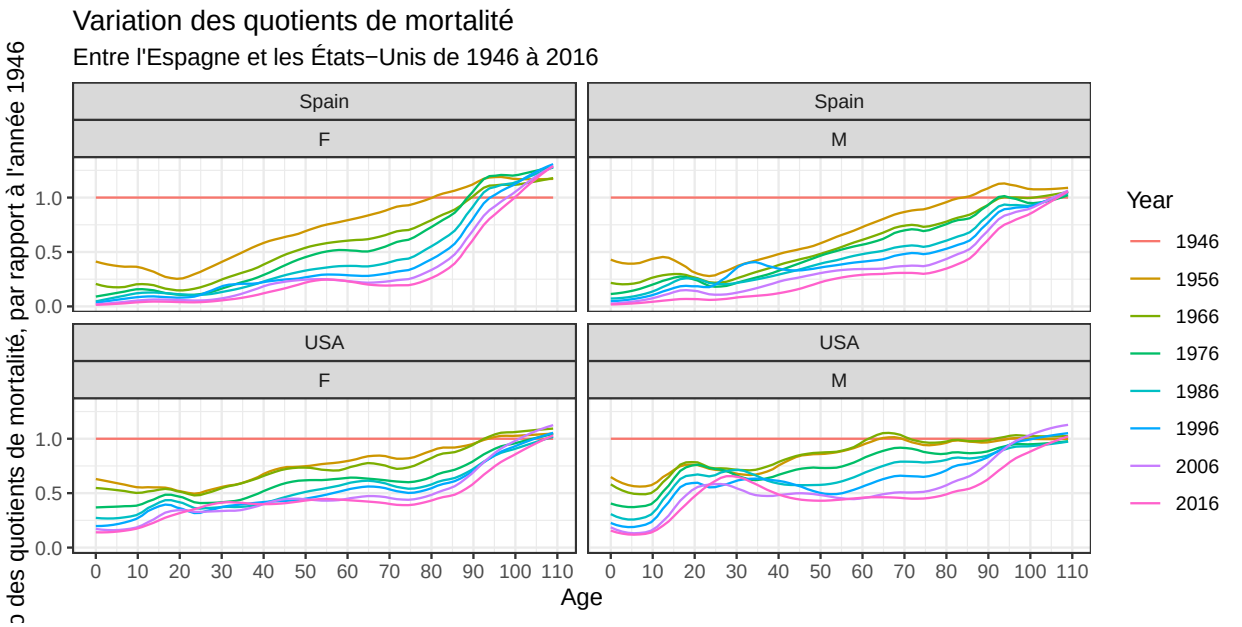
- Peu importe le sexe, à l'âge adulte (c'est à dire à partir de 23 ans), nous observons une forte diminution des ratios des quotients de mortalité entre les deux pays atteignant un minimum globale à l'âge de 57 ans. A cet âge, le risque annuel de décès pour les filles est 0.85 supérieur à celui des États-Unis. Concernant les garçons, ce risque annuel de décès est supérieur à 0.84 à celui des États-Unis.

- A partir de 60 ans, les ratios des quotients de mortalité se stabilisent nous montrant que le risque annuel de décès à ces âges élevés est très proche entre l'Espagne et les États-Unis.

Évolution des quotients de mortalité depuis la Seconde Guerre mondiale



- Concernant le graphique ci-dessus, nous constatons que les quotients de mortalité des jeunes en 1946 est plus petit aux Etats-Unis par rapport à l'Espagne. Cela est certainement dû au fait que les États-Unis n'ont pas beaucoup souffert de pertes humaines pendant la Seconde Guerre mondiale, contrairement à l'Espagne.

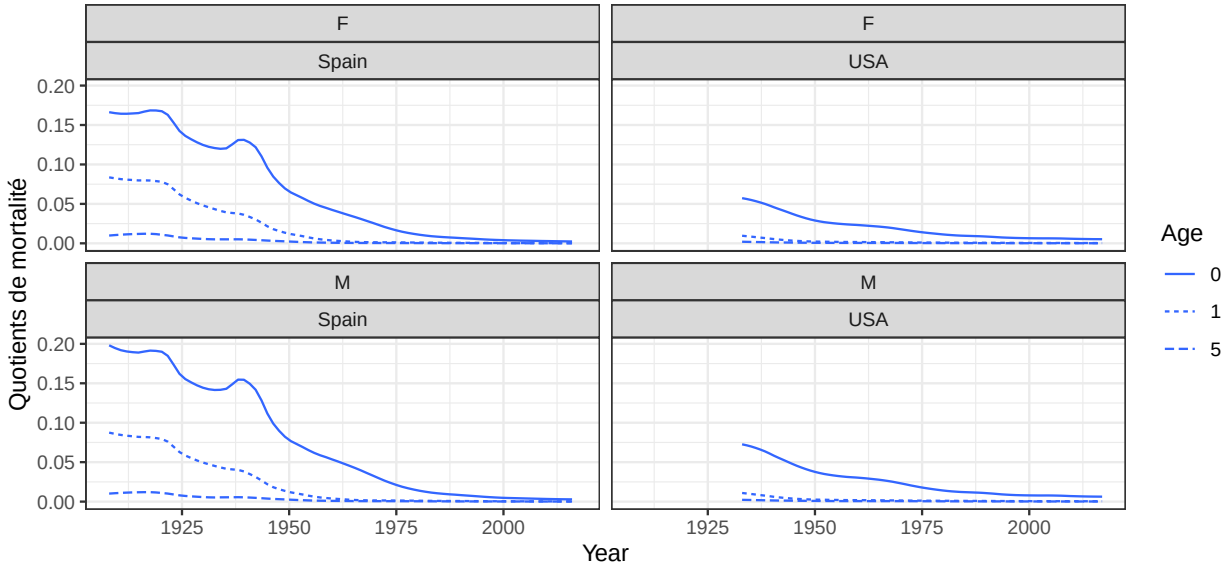


- Ce graphique nous informe que, pour les deux pays, les quotients de mortalité diminuent au fil des années pour les âges <= 80 ans.
- En effet, en 1956 pour l'Espagne , le ratio est de 0.5 à la naissance et celui-ci passe à 0.25 en 1966.
- Nous constatons une diminution de la mortalité plus marqué en Espagne par rapport aux Etats-Unis notamment chez les hommes.

- Nous observons que le risque de mortalité pour les personnes âgées (les femmes), notamment en Espagne est plus élevé par les années 1956 à 2016 par rapport à l'année 1946.

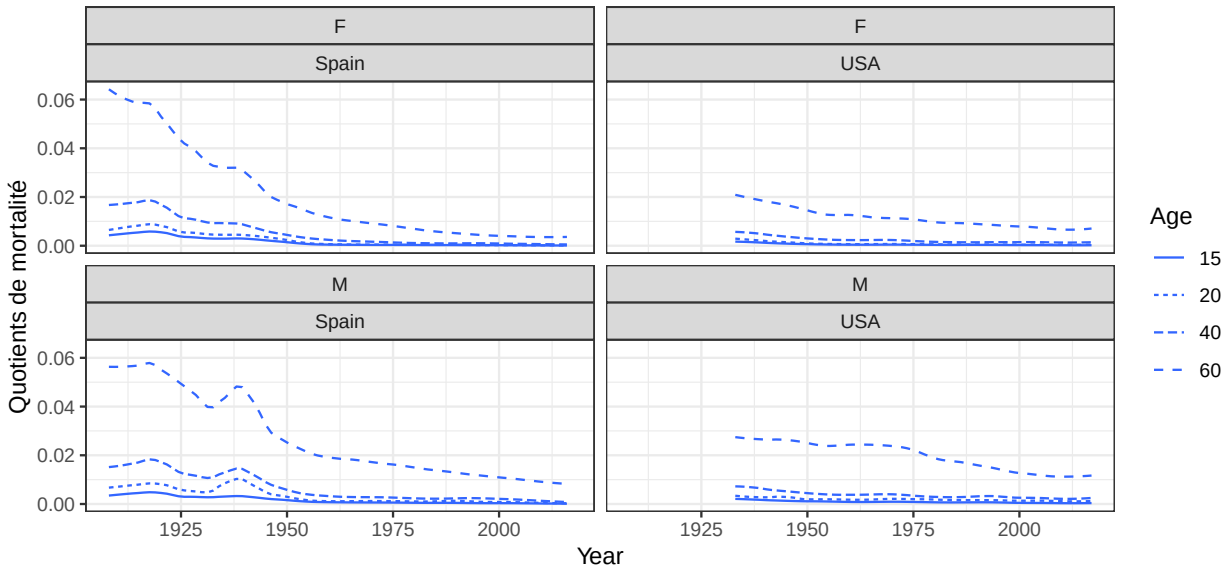
Tendances

Evolution des quotients de mortalité pour les nouveaux nés et les enfants.



- Ce graphique nous informe que les taux de mortalité les plus élevés sont à la naissance quelque soit le sexe et la période donnée.
- Les taux de mortalité à la naissance/infantile ont fortement diminué au cours du temps en Espagne et aux États-Unis. Les quotients de mortalité des âges $0, 1, 5$ sont très différents dans le siècle précédent, pourtant ils se rapprochent à partir de 2000.

Evolution des quotients de mortalité pour les adolescents et les adultes.



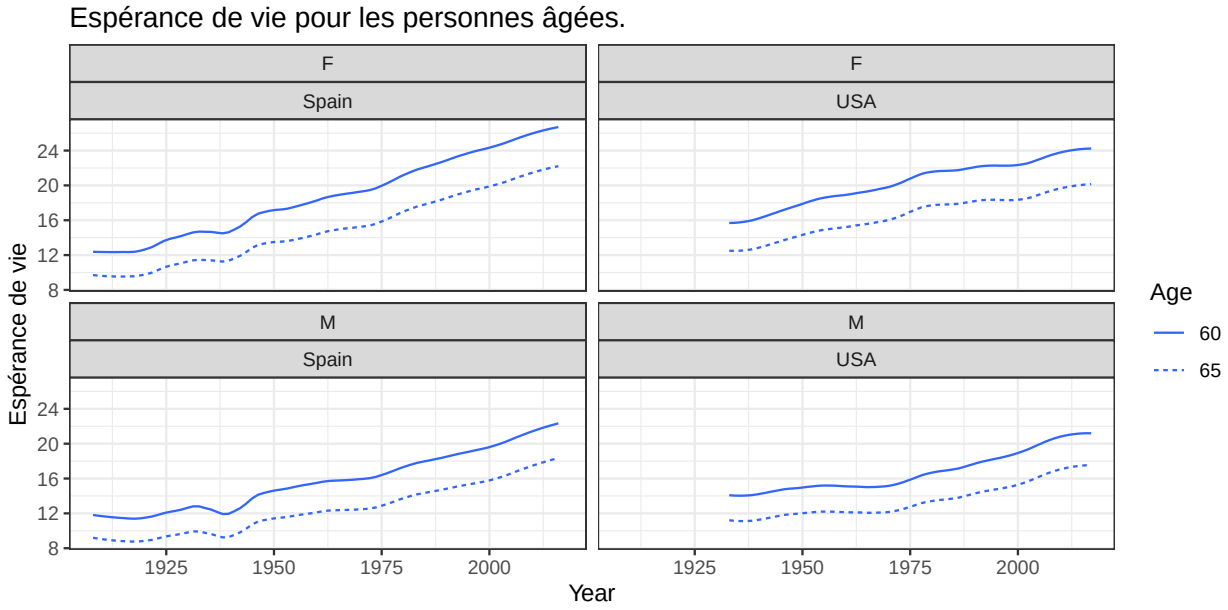
- Ce graphique nous informe que les taux de mortalité les plus élevés sont aux âges les plus élevés (i.e 60 ans) quelque soit le sexe et la période donnée.
- Les taux de mortalité aux âges élevés ont diminué au cours du temps en Espagne et aux États-Unis. Signifiant que l'espérance de vie à la naissance pour ces âges élevés a augmentée durant les années 2000.
- Nous constatons que les taux de mortalité pour les adolescents âgés de 15 ans est relativement constante à travers le temps.

Rearrangement des données

Nous réarrangeons nos données en pivotant notre table universelle de départ par rapport à la colonne Age. Chaque ligne est identifiée par (Year,Gender,Country). Les colonnes correspondent aux âges et le contenu de chaque cellule est le log-quotient de mortalité à un âge donné pour une année, un pays et un sexe donné.

```
## # A tibble: 2,298 x 113
##     Year Gender Country   `0`    `1`    `2`    `3`    `4`    `5`    `6`    `7`
##    <int> <fct>  <fct>   <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
##  1  1816 F      France  0.187 0.0467 0.0339 0.0229 0.0160 0.0138 0.0121 0.0104
##  2  1817 F      France  0.182 0.0542 0.0389 0.0270 0.0188 0.0152 0.0129 0.0107
##  3  1818 F      France  0.186 0.0610 0.0417 0.0286 0.0203 0.0165 0.0140 0.0118
##  4  1819 F      France  0.197 0.0662 0.0456 0.0300 0.0215 0.0181 0.0158 0.0136
##  5  1820 F      France  0.181 0.0561 0.0393 0.0262 0.0184 0.0158 0.0140 0.0123
##  6  1821 F      France  0.182 0.0567 0.0414 0.0280 0.0195 0.0162 0.0142 0.0121
##  7  1822 F      France  0.207 0.0602 0.0421 0.0291 0.0202 0.0164 0.0140 0.0118
##  8  1823 F      France  0.192 0.0556 0.0374 0.0251 0.0181 0.0153 0.0134 0.0118
##  9  1824 F      France  0.199 0.0616 0.0433 0.0281 0.0194 0.0159 0.0134 0.0112
## 10  1825 F      France  0.194 0.0637 0.0450 0.0303 0.0206 0.0168 0.0143 0.0121
## # ... with 2,288 more rows, and 102 more variables: `8` <dbl>, `9` <dbl>,
## #   `10` <dbl>, `11` <dbl>, `12` <dbl>, `13` <dbl>, `14` <dbl>, `15` <dbl>,
## #   `16` <dbl>, `17` <dbl>, `18` <dbl>, `19` <dbl>, `20` <dbl>, `21` <dbl>,
## #   `22` <dbl>, `23` <dbl>, `24` <dbl>, `25` <dbl>, `26` <dbl>, `27` <dbl>,
## #   `28` <dbl>, `29` <dbl>, `30` <dbl>, `31` <dbl>, `32` <dbl>, `33` <dbl>,
## #   `34` <dbl>, `35` <dbl>, `36` <dbl>, `37` <dbl>, `38` <dbl>, `39` <dbl>,
## #   `40` <dbl>, `41` <dbl>, `42` <dbl>, `43` <dbl>, `44` <dbl>, `45` <dbl>,
## #   `46` <dbl>, `47` <dbl>, `48` <dbl>, `49` <dbl>, `50` <dbl>, `51` <dbl>,
## #   `52` <dbl>, `53` <dbl>, `54` <dbl>, `55` <dbl>, `56` <dbl>, `57` <dbl>,
## #   `58` <dbl>, `59` <dbl>, `60` <dbl>, `61` <dbl>, `62` <dbl>, `63` <dbl>,
## #   `64` <dbl>, `65` <dbl>, `66` <dbl>, `67` <dbl>, `68` <dbl>, `69` <dbl>,
## #   `70` <dbl>, `71` <dbl>, `72` <dbl>, `73` <dbl>, `74` <dbl>, `75` <dbl>,
## #   `76` <dbl>, `77` <dbl>, `78` <dbl>, `79` <dbl>, `80` <dbl>, `81` <dbl>,
## #   `82` <dbl>, `83` <dbl>, `84` <dbl>, `85` <dbl>, `86` <dbl>, `87` <dbl>,
## #   `88` <dbl>, `89` <dbl>, `90` <dbl>, `91` <dbl>, `92` <dbl>, `93` <dbl>,
## #   `94` <dbl>, `95` <dbl>, `96` <dbl>, `97` <dbl>, `98` <dbl>, `99` <dbl>,
## #   `100` <dbl>, `101` <dbl>, `102` <dbl>, `103` <dbl>, `104` <dbl>,
## #   `105` <dbl>, `106` <dbl>, `107` <dbl>, ...
```

Espérance de vie



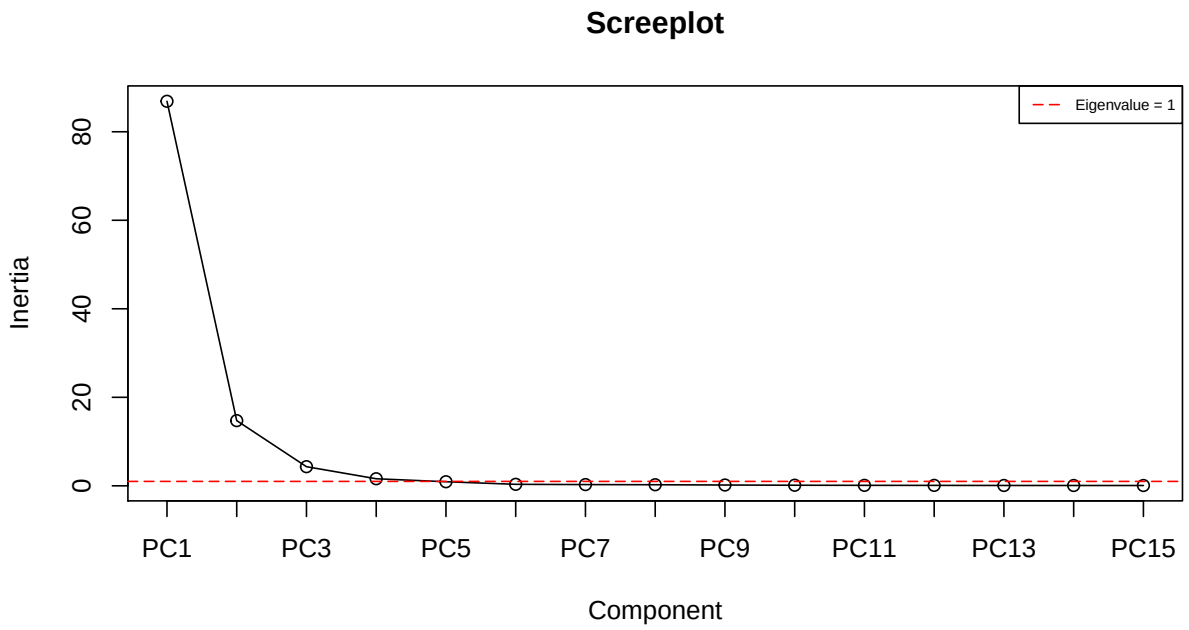
- Nous constatons que l'espérance de vie chez les personnes âgées $(60, 65$ ans) ont progressivement augmenté du siècle précédent jusqu'à présent, par exemple : une femme espagnole qui pouvait espérer de vivre encore 12 ans en 1900, peut en 2000, espérer deux fois plus longtemps soit près de 24 ans.

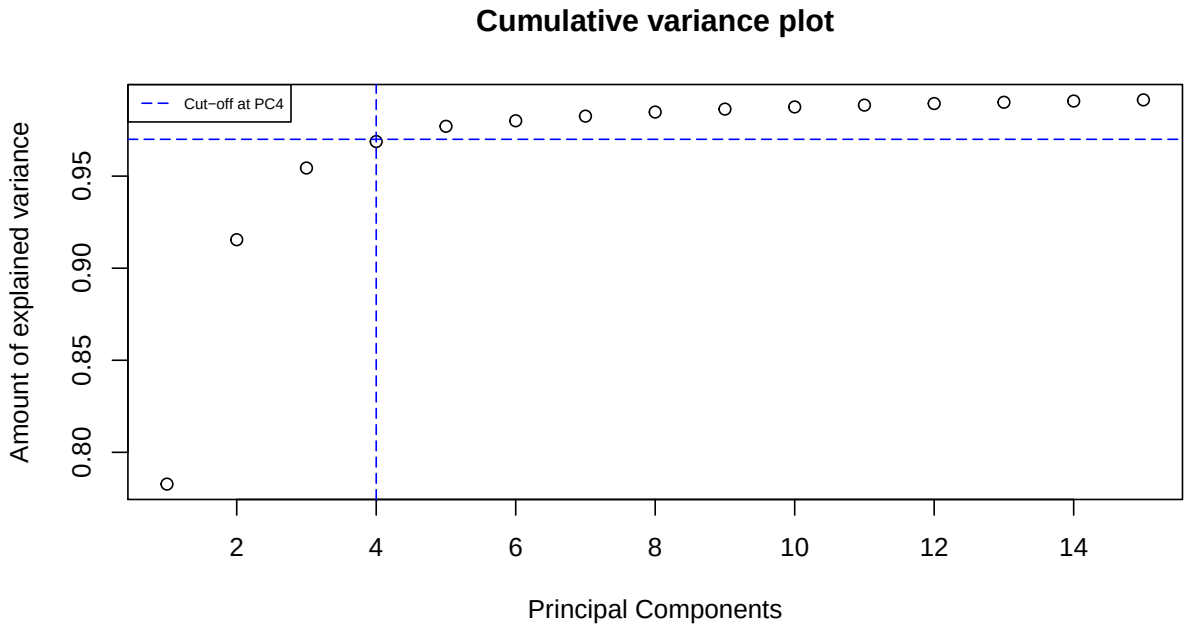
Applications de méthodes factorielles sur données

PCA et SVD sur les tables de log-mortalité

L'analyse des composantes principales (ACP , i.e PCA) est une technique de réduction de dimensions. Notre table de données df_pivot comporte un grand nombre de variables $X_0, ..., X_{109}$ dont certaines sont corrélées. Cette corrélation entre les variables entraîne une redondance de l'information au sein de nos données. Etant que nous avons 110 variables,nous ne pouvons pas obtenir une bonne visualsation en raison du grand nombre de dimensions.Nous utilisons une PCA pour transformer les variables originales $X_0, ..., X_{109}$ en une combinaison linéaire de ces variables qui sont indépendante. A l'issue de cette réduction de dimension, il est dorénavant possible de représenter dans un graphique à deux dimensions, les variables originales par rapport aux nouvelles variables (les axes de notre graphique).

Exemple d'utilisation de PCA : Hommes - Espagne.



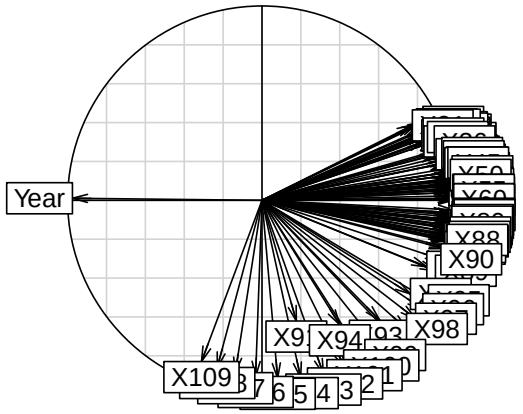


Le screeplot nous montre la quantité de variation que chaque composante principale capture à partir des données.

La règle de Kaiser (on ne considère que les axes associés à une valeur propre strictement supérieure à 1) nous assure que nous devons garder les quatre premières composantes principales afin de garder un maximum d'inertie (information) pour un nombre minimale de composantes.

Choisir les quatre premières composantes principales (Cut-off à PCA 4) permet d'expliquer 96% des variations initiales.

Le screeplot est une courbe raide qui se courbe rapidement puis s'aplatit au point de "Cut-off".



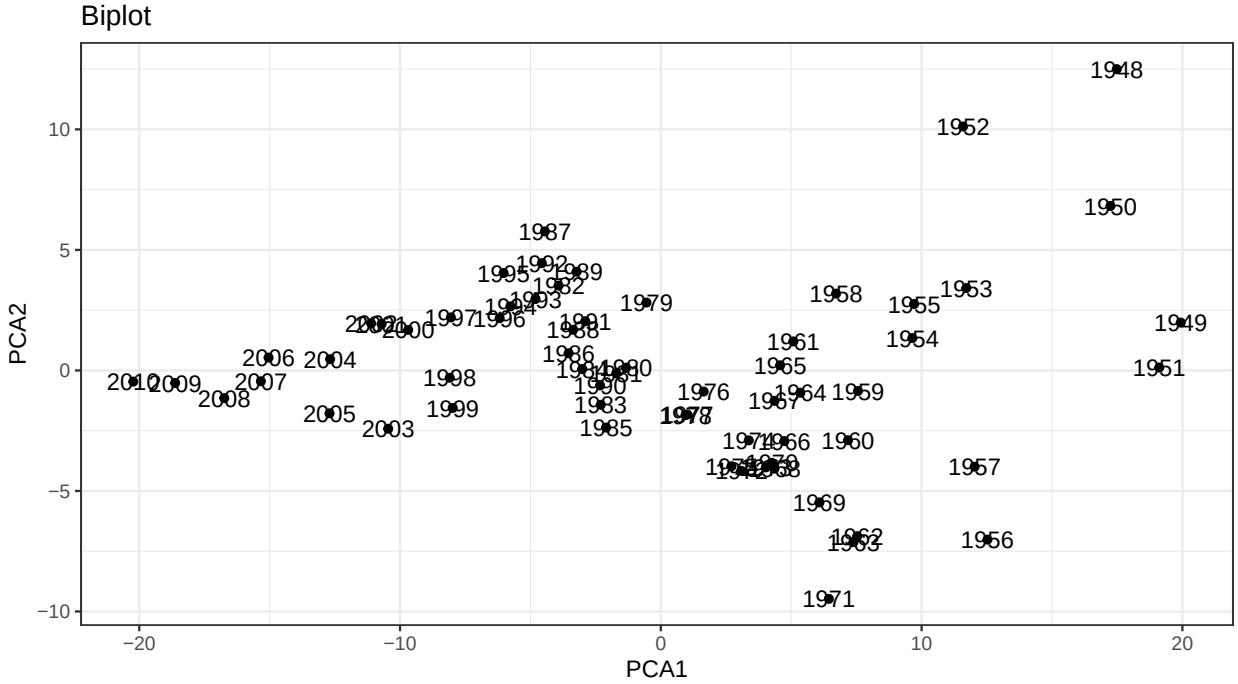
Ce graphique représente les corrélations entre les variables âges $X_0, ..., X_{109}$. Il montre les relations entre toutes ces variables.

Nous notons que les vecteurs X_0 à X_{90} ont la même direction. Ces vecteurs représentant les variables d'âge sont corrélés positivement. L'angle entre deux flèches dans ce groupe de vecteurs est très petit, ce qui montre que ces variables sont très positivement corrélées.

Les vecteurs X_{91} à X_{109} ont la même direction. Montrant qu'il y a une corrélation entre ces variables. Cependant, l'angle entre deux flèches dans ce groupe de vecteurs n'est pas assez petit pour dire qu'ils sont très positivement corrélés comme avant.

Notons que les vecteurs X_{105} et X_{60} sont perpendiculaires, les variables sont non corrélées (indépendantes).

Nous constatons que les vecteurs X_{109} et X_{24} divergent l'un de l'autre et forment un angle de 180ř. Ces vecteurs sont corrélés négativement.



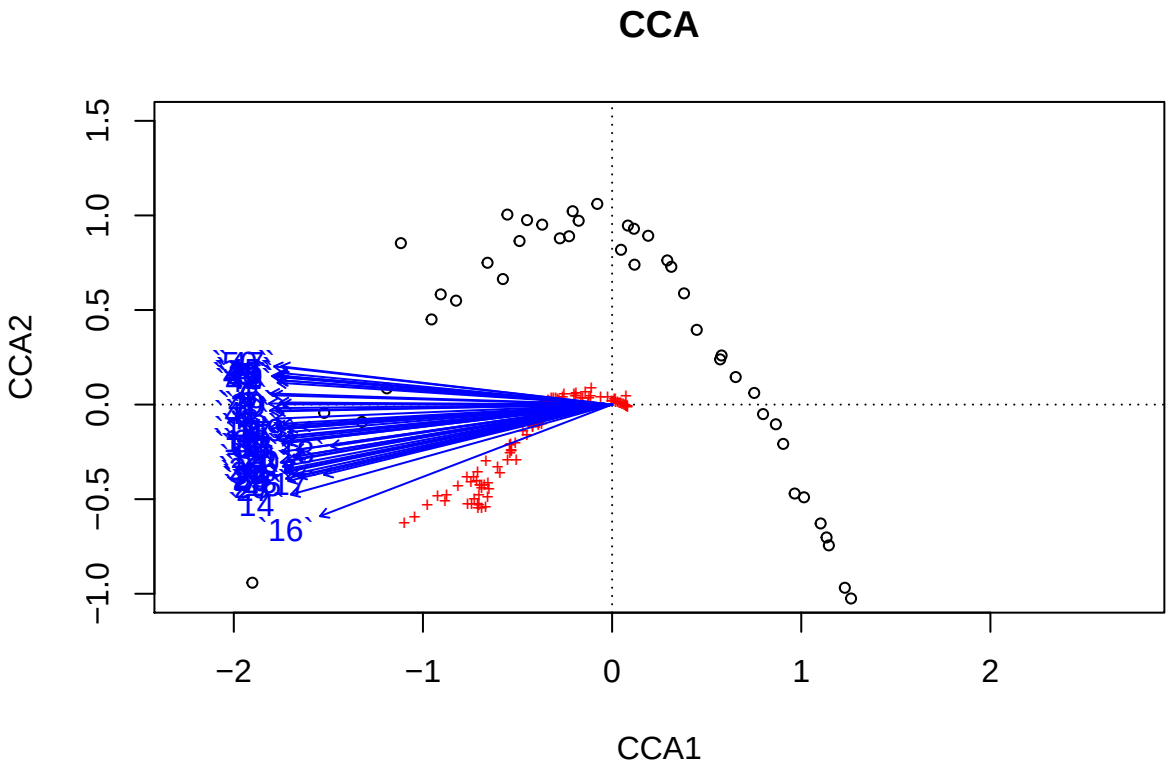
Le biplot consiste simplement à fusionner le cercle de corrélation avec le graphique des individus (représentant le positionnement des années par rapport aux composantes principales).

Nous préférons représenter les deux graphiques séparément, par défaut de la lisibilité. Néanmoins, une analyse pertinente est possible :

Les facteurs les plus importants pour PC2 sont X_{99} à X_{109}. Leurs effets vont dans la même direction. Ces variables, formant un cluster d'âge élevé, tendent à nous faire comprendre qu'un taux de mortalité à ces âges est associé aux années 1980 à 2010.

De nombreux facteurs contribuent à PC1, dont X_0 à X_{90} et leurs effets vont également dans la même direction. Ces facteurs, formant une cluster d'âge, tendent à nous faire comprendre qu'un taux de mortalité a ces âges est associé pour les années 1950 à 1980.

Analyse correspondance canonique



Interpretation :

- Les longueurs et positions des flèches fournissent des informations sur la relation entre les variables d'âges et les axes. Les flèches parallèles à un axe (par exemple la variable '16' et l'axe CCA1) indiquent une corrélation, la longueur de la flèche nous indique la force de cette corrélation. Ainsi, l'ensemble de nos variables âges $1, ..., 109$ sont fortement liées à l'axe CCA1 mais aucun de ces éléments n'est lié à l'axe CCA2.
- L'axe CCA1 est associé à l'ensemble de ces variables d'âges, cependant puisque la corrélation entre ces variables et l'axe CCA1 est négative, les scores élevés positifs sur l'axe CCA1 devraient avoir de faible valeurs pour la variable '16', tandis que les scores élevés négatifs sur l'axe CCA1 devraient avoir une très forte corrélation avec nos variables de départ.

- Ainsi, nous nous attendrions à ce que l'ensemble croix rouges soient fortement corrélés à ces variables tandis que la moitié des points blancs ait peu de relations avec ces variables.

Modèle de Lee-Carter et ses limites

Présentation du modèle

Au cours du siècle dernier, aux États-Unis et en Europe occidentale, les quotients de mortalité à tous les âges ont affiché une tendance générale à la baisse comme nous l'avons pu constaté précédement. Cette tendance à la baisse n'a pas toujours été homogène d'un âge à l'autre.

Le modèle Lee-Carter a été conçu pour modéliser et prévoir l'évolution des log-quotients de mortalité aux États-Unis au cours du XXe siècle.

Soit $log(m_{x,t})$ le log-quotient de mortalité à l'âge x au cours de l'année $t \in T$ pour une population donnée (définie par le sexe et le pays).

Le modèle de Lee-Carter suppose que les log-quotient de mortalité observés sont échantillonnés selon le modèle suivant

$$log(m_{x,t}) \sim_{\text{indépendant}} a_x + b_x\kappa_t + \epsilon_{x,t}$$

où $(a_x)_x, (b_x)_x$ et $(\kappa_t)_t$ sont des vecteurs inconnus avec :

$$a_x = \frac{1}{|T|}\sum_{t\in T} log(m_{x,t}) \qquad \sum_{t\in T}\kappa_t = 0 \qquad \sum_x b_x^2 = 1$$

et $\epsilon_{x,t}$ des variables aléatoires gaussiennes i.i.d.

Le paramètre $(a_x)_x$ s'interprète comme la valeur moyenne des $log(m_{x,t})$ au cours du temps.

Le paramètre $(b_x)_x$ traduit la sensibilité de la mortalité instantanée à l'âge x par rapport à l'évolution générale de l'indice de mortalité k_t.

$\epsilon_{x,t}$ sont les termes d'erreurs qui reflètent les influences historique résiduelle spécifique à l'âge non captées par le modèle.

On obtient les paramètres par un critère de moindres carrés :

$$(\hat{a}_x, \hat{b}_x, \hat{k}_t) = argmin \sum_{x,t}(ln(m_x) - a_x - b_x k_t)^2$$

Cette estimation des moindres carrées nous donne la plus petite valeur pour la somme des erreurs au carré. Pour trouver ces coefficients (s'appelant la phase de "fitting" ou d'apprentissage du modèle) les auteurs emploient une décomposition en valeurs singulières (i.e SVD) pour trouver la solution à ce problème de minimisation.

Considérons la matrice Z de dimension (111,n-d) où d et n sont respectivement les dates de début (d) et date de fin (n) de colonne la Year pour un pays donné :

$$Z = (log(m_{x,t}) - \hat{a}_x)_{x,t})$$

En appliquant une SVD à la matrice Z, on obtient la décomposition suivante :

$$SVD(Z) = PDQ$$

$$SVD(Z) = D_1 P_{x1} Q_{t1} + D_2 P_{x2} Q_{t2} + ... + D_w P_{x_w} Q_{t_w}$$

où rank(Z) = w , D_i où $i = 1, ..., w$ correspond aux valeurs singulières classées par ordre décroissant, P_{x_i} et $Q_{t,i}$ où $i = 1, ..., w$ correspondant conformément aux vecteurs singuliers gauche et droit.

Dans le cas du modèle de Lee-Carter, les auteurs effectuent une approximation pour les premiers termes, en considérant les termes suivants :

$$\hat{b}_x = P_{x_1}$$
$$\hat{k}_x = D_1 Q_{t_1}$$

L'une des propriétés remarquables du modèle de Lee-Carter est qu'après cette phase de "fitting" (c'est-à-dire une fois que les valeurs de a_x, b_x et k_t sont calculées), l'étape de "forecasting" des quotients de mortalité se réduit à prédire l'indice de mortalité k_t comme nous le verrons dans la suite.

Limites & Ouverture

L'une des principales limite à ce modèle est que les termes d'erreurs sont supposées être homoscédastique et de distribution normale. En d'autres termes, cela revient à dire que :

$$\text{Var}[\epsilon_{x,t}] = \sigma_i \forall i$$

Cette hypothèse est irréaliste pour modéliser la mortalité humaine. Les log-quotients de mortalité observés sont beaucoup plus variable aux personnes âgés qu'aux âges plus jeunes. La variance des taux de décès croit aux âges plus élevées, du fait notamment de la baisse des effectifs des survivants.

Le critère de sélection des paramètres optimaux n'a pas de justification probabiliste. Nous nous sommes poser les questions suivantes : peux-on estimer les paramètres autrement ? Est-ce une bonne ou mauvaise idée ? Si oui, de quelle manière ?. Chercher à modéliser l'approche pour obtenir un critère de détermination des paramètres grâce à des méthodes de type "maximum de vraisemblance" peut-être intéressant (cet estimateur existe-il ? peut-on le calculer ? si oui, le calculer et voir que ce n'est pas une bonne idée.). En effet, nous pourrions bénéficier de bonnes propriétés de cette classe d'estimateur (convergence, efficacité asymptotique, normalité asymptotique etc ..).

Application du modèle de Lee-Carter sur nos données

Dans cette partie, on s'intéresse à l'application du modèle Lee-Carter pour prédire les quotients de mortalité et les espérances de vie de tous les âges.

Lee et Carter suppose que $log(m_{x,t}) = a_x + b_x k_t + e_{x,t}$ ce qui est équivaut à dire que $m_{x,t} = exp(a_x + b_x k_t + e_{x,t})$. Pour simplifier, comme $e_{x,t}$ est le terme d'erreurs qui a un impact négligeable sur la valeur de $m_{x,t}$ et est difficile à prédire, on suppose que : $m_{x,t} = exp(a_x + b_x k_t)$. Pour ce fait, nous utilisons les données historiques, disponibles pour calculer les coefficients a_x, b_x et k_t où x correspond à l'âge et t correspond à l'année.

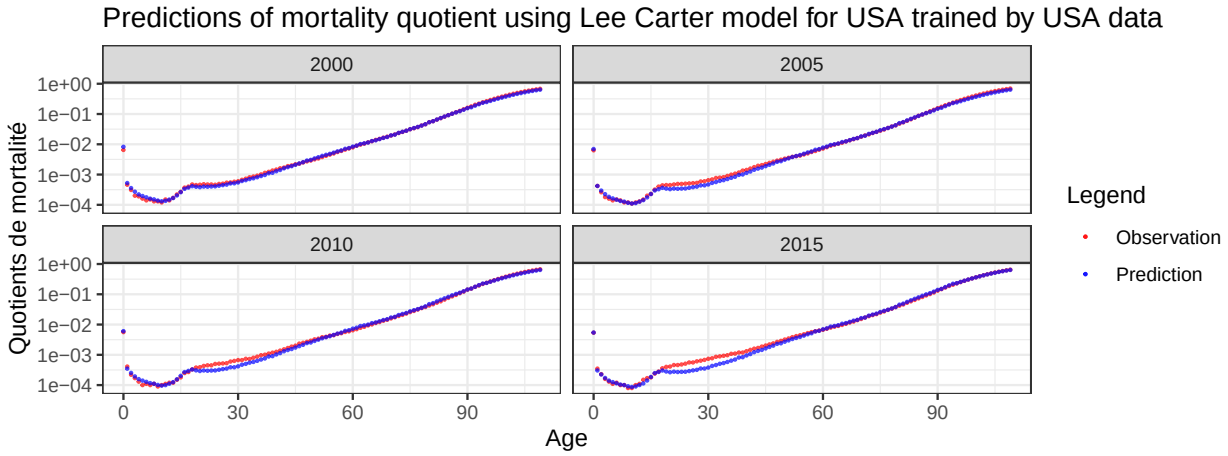
Dans un premier temps, nous allons travailler avec les données américaines. Nous allons entraîner notre modèle avec les données historiques de 1933 à 1995 et essayer de prédire les quotients de mortalité des États-Unis pour les années 2000, 2005 , ...,2015. Ensuite, nous allons comparer les prédictions avec les observations réelles pour vérifier la validation du modèle.

Comme a_x et b_x sont des variables indépendantes du temps, nous pouvons facilement les calculer avec les données historiques (cf la partie précédente - critère de moindres carrés - SVD) et les réutiliser pour la prédiction. Or k_t est une variable dépendante du temps. Ainsi, il faut une méthode pour estimer ces variables afin de pouvoir réaliser une prédiction des quotients de mortalité.

Appelons k une fonction de $\mathbb{N}$ dans $\mathbb{R}$ qui prends une année t en argument et donne la valeur k_t. On peut simuler k par un processus stochastique, comme une marche aléatoire. On a ainsi $k_t = k_{t-1} + d + e_t$ avec $d = \frac{k_{LastYear} - k_{FirstYear}}{LastYear - FirstYear}$ étant le "drift" et e_t le terme d'erreur suivant une loi normale de paramètre $(0, \nu)$ avec ν la moyenne des variances des k_{t_obs}.

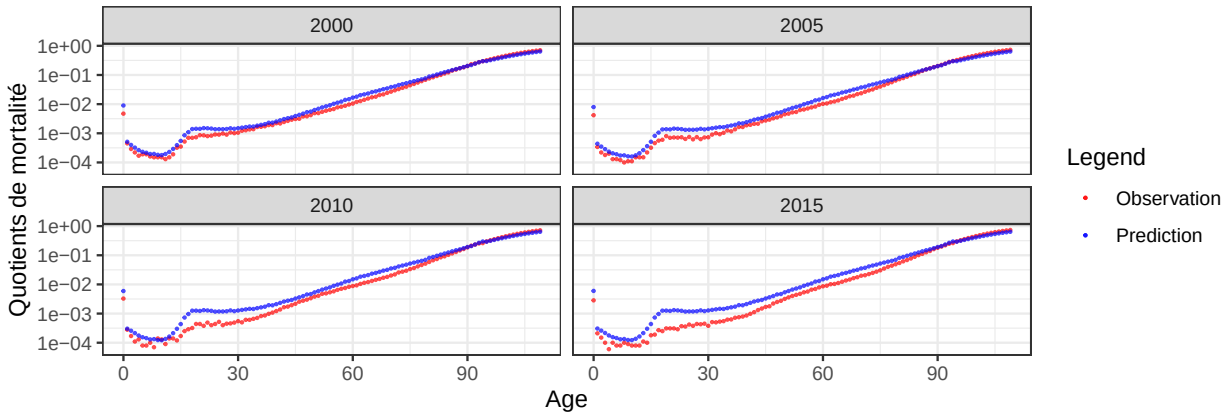
En utilisant les prédictions de k_t données par ce processus stochastique et les paramètres a_x et b_x obtenus grâce une estimation des moindres carrées, nous les utilisons dans la formule suivante $m_{x,t} = exp(a_x + b_x k_t)$. Ainsi, nous pouvons calculer et prédire les quotients de mortalité pour les années à venir.

Appliquons ce procédé aux données américaines (femmes). Notre phase de "fitting" i.e d'apprentissage concernera les années $1933, ..., 1995$ afin de pouvoir faire du "forecasting" i.e des prédictions pour les années $2000, 2005, ..., 2015$. Nous comparerons les prédictions obtenues et les observations de ces années :

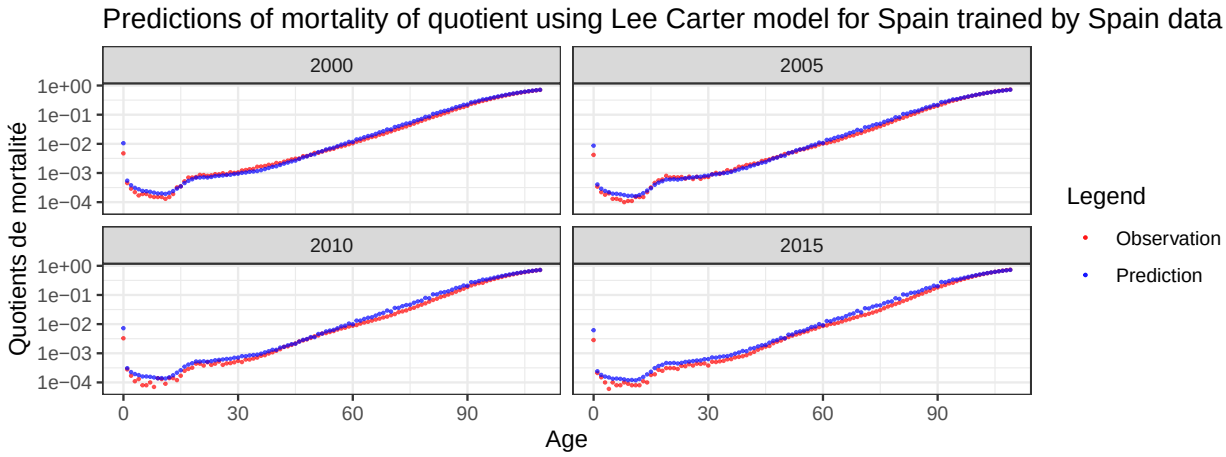


Nous constatons que le modèle de Lee-Carter donne des prédictions assez proches de la réalité, toutefois en sous-estimant le quotient de mortalité.

Essayons maintenant d'utiliser les données américaines (Hommes) pour prédire les quotients de mortalité chez les hommes espagnols :

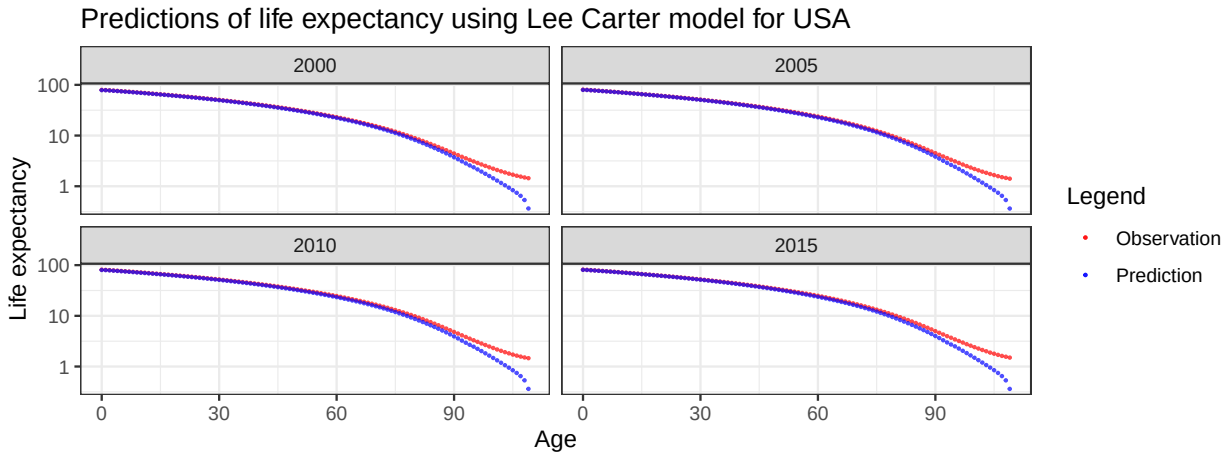


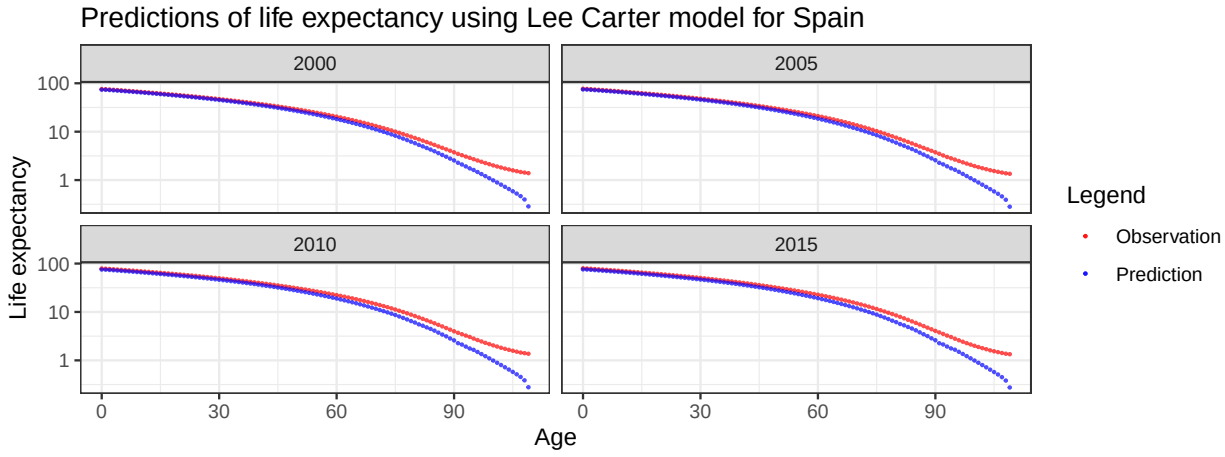
Nous pouvons directement observer que le modèle marche moins bien pour l'Espagne que les Etat-Unis. En revanche, utilisant les données américaines pour prédire les quotients de mortalité espagnols ne semble pas très logique parce que chaque pays est différent des autres avec ses propres caractéristiques. Essayons maintenant d'utiliser les données espagnoles:



Nous constatons ainsi une prédiction très fiable et correspondante à la réalité, ce qui prouve la validité du modèle de Lee-Carter.

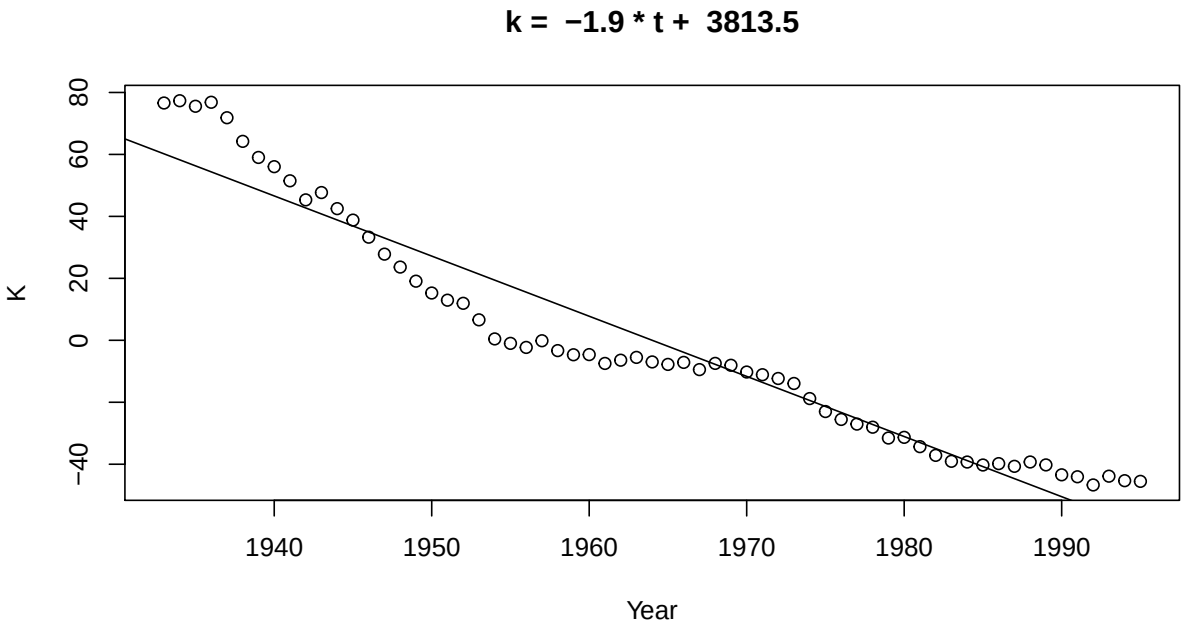
Avec les quotients de mortalité prédits, nous pouvons faire une prédiction de l'espérance de vie comme ci-dessous:



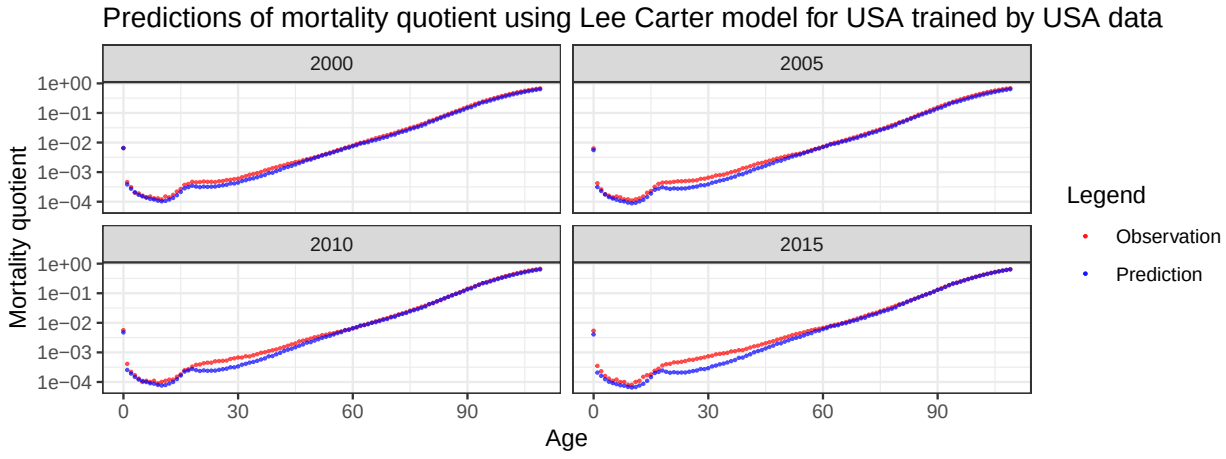


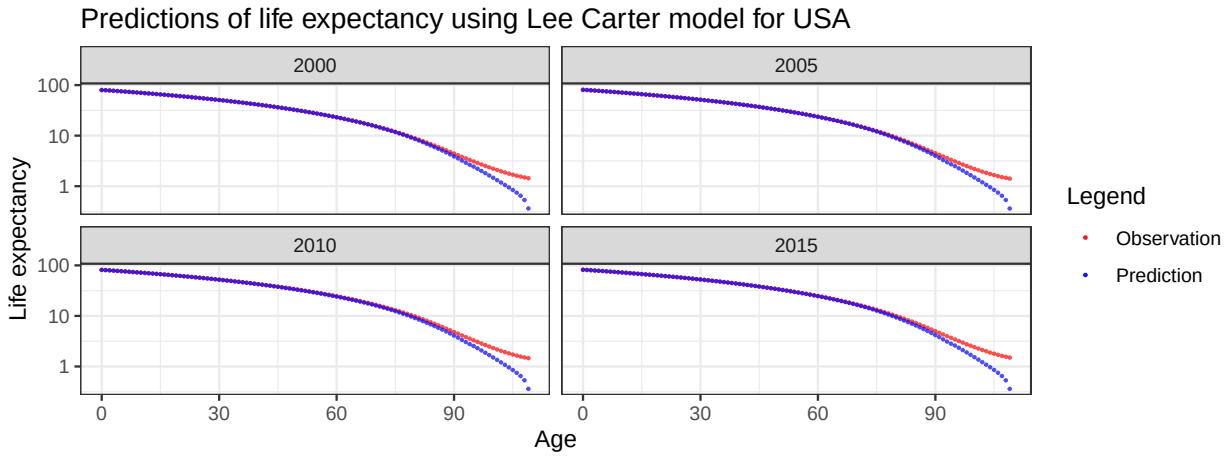
Nous constatons que le modèle Lee-Carter donne des prédictions fiables pour les jeunes (avant l'âge 60), pourtant pour les âges élevés (à partir de 60 ans), le modèle Lee-Carter semble un peu pessimiste, en sous-estimant les espérances de vie chez les personnes âgées.

Une autre méthode pour estimer k_t, différente de la méthode précédente, est de faire une régression linéaire. En effet, k_t est décroissante presque linéairement en fonction du temps. Cette décroissance linéaire de k est à la base de la méthode de Lee-Carter pour prévoir la mortalité. Nous pouvons appliquer une régression linéaire pour les données américaines de 1933 à 1995 (Femmes) comme ci-dessous:



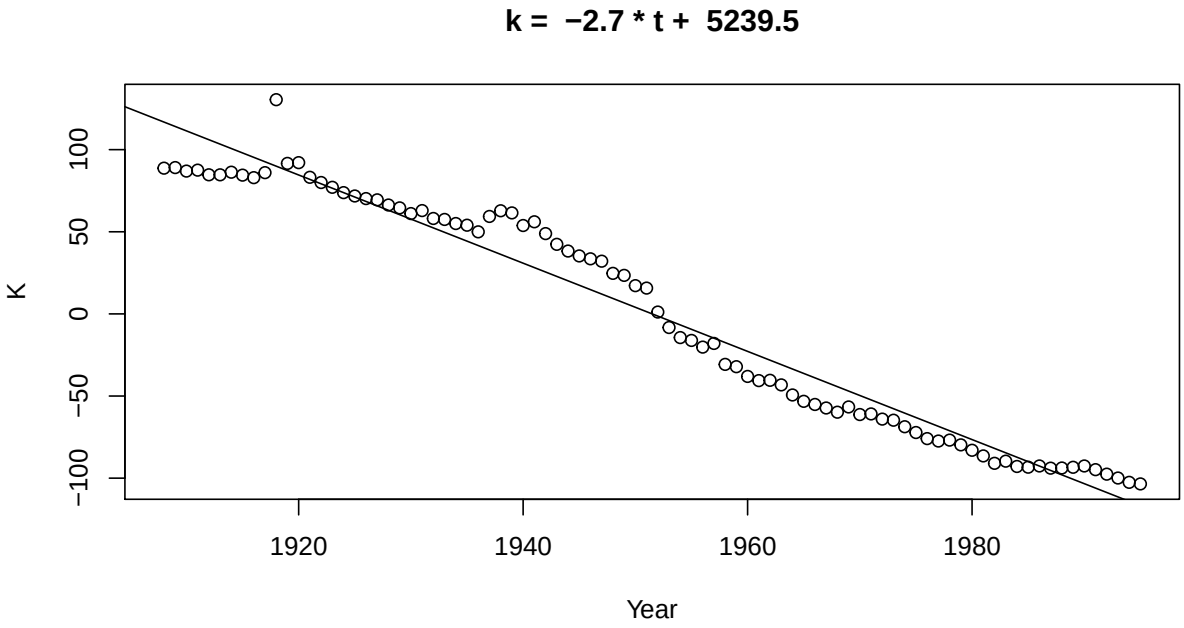
En utilisant les prédictions de k_t données par la régression linéaire et les paramètres a_x et b_x obtenus grâce une estimation des moindres carrées. Nous pouvons les utiliser dans la formule suivante $m_{x,t} = exp(a_x + b_x k_t)$. Ainsi, nous pouvons calculer et prédire les quotients de mortalité pour les années à venir. Nous obtenons ainsi les prédictions suivantes:





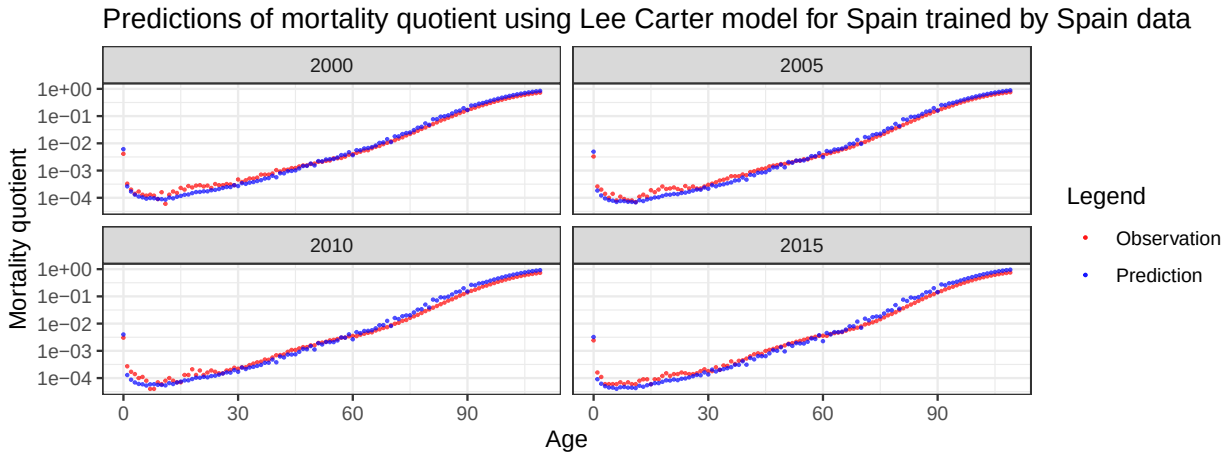
A titre de remarque, nous avons pu constaté que le modèle de Lee-Carter avec une régression linéaire fonctionne aussi bien que celui présenté précédement. Il donne toujours des prédictions fiables des quotients de mortalité et des espérances de vie.

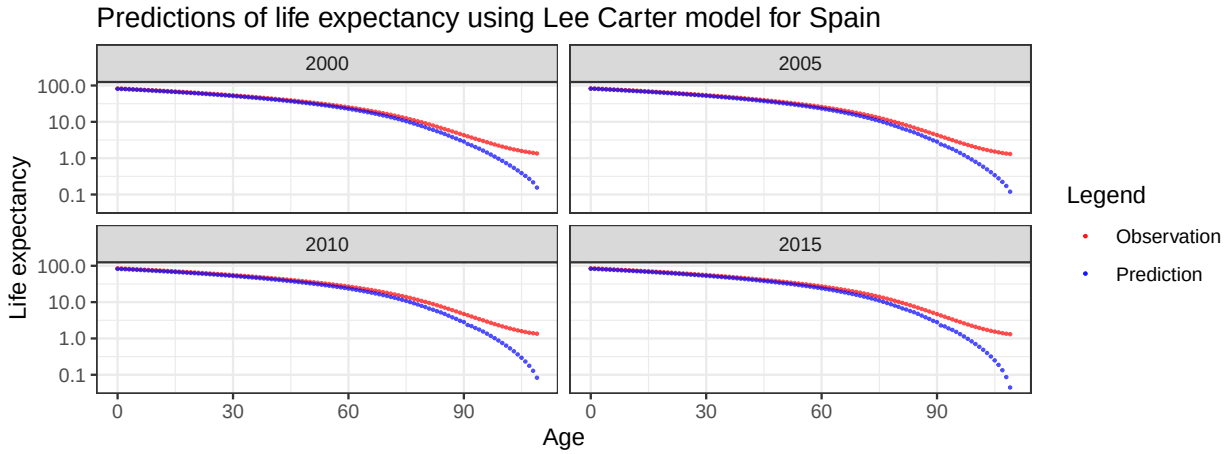
Nous pouvons maintenant essayer de voir si le déclin de k est linéaire pour autre pays ou s'il est le cas seulement pour les Etats-Unis. Appliquons une régression linéaire pour les k_t de l'Espagne entre 1900 et 1995 (Femmes):



Nous constatons ici que la régression linéaire reste très valide pour les données espagnoles de 1900 à 1995. Nous observons aussi que la pente de k_{Spain} est supérieure à celle de k_{USA}, cela a pour conséquent : les quotients de mortalité de l'Espagne vont diminuer plus rapide que ceux d'USA (nous avons constaté ce phénomène dans le graphique "Variation des quotients de mortalité entre Espagne et Etats-Unis")

Avec les prédictions de cette régression linéaire, nous pouvons espérer des prédictions fiables pour les années 2000 à 2015. Appliquons cela aux années 2000 à 2015 afin de valider notre hypothèses:





Nous constatons finalement une prédiction très fiable pour les femmes espagnoles avec le modèle Lee-Carter-Régression linéaire.