Phase 2 du Projet Fil Rouge :

Analyse exploratoire de la base de données « MortalInval »

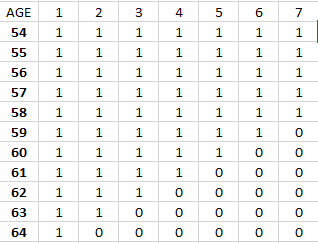
# Présentation du jeu de données MortalInval

La base de données soumise à notre étude est constituée de trois tableaux de données disjoints décrivant la dynamique de la population d’une classe d’âge connue sur toute la durée de leur vie active au sein d’une entreprise.

Le premier tableau présente le recense ment de la population de classes d’âge allant de 25 à 64 ans à l’instant to sur toute la période de leur activité qui s’achève à 65ans.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Age** | **0** | **1** | **2** | **3** | **4** | **5** |
| **25** | 10000 | 9953 | 9854 | 9792 | 9765 | 9706 |
| **26** | 10000 | 9938 | 9837 | 9784 | 9718 | 9669 |
| **27** | 10000 | 9956 | 9837 | 9800 | 9731 | 9650 |
| **28** | 10000 | 9918 | 9862 | 9762 | 9709 | 9641 |
| **29** | 10000 | 9913 | 9812 | 9762 | 9700 | 9595 |
| **30** | 10000 | 9912 | 9805 | 9764 | 9666 | 9591 |

Le tableau 2 présente pour chaque groupe d’individus de classe d’âge donné l’indicateur de statuts valant 0 ou 1 selon que l’observation correspond à une censure ou non sur toute la durée d’observation.



Le dernier tableau présente le nombre d’individus qui sont soties de l’entreprise à chaque date d’observations pour les différentes classes d’âge observée.

# Analyse exploratoire de la base de données

Dans cette analyse de données, nous nous focalisons sur les données brutes issues du premier tableau de données « MortalInval ».

Nettoyage des données

A partir de la base de données initiale, nous réalisons une opération borélienne de transposition et rajoutons une colonne supplémentaire représentant la date d’observation de sorte à constituer une table de survie. Nous procédons également à la nomenclature des observables ainsi que des objets.

## Visualisation des données

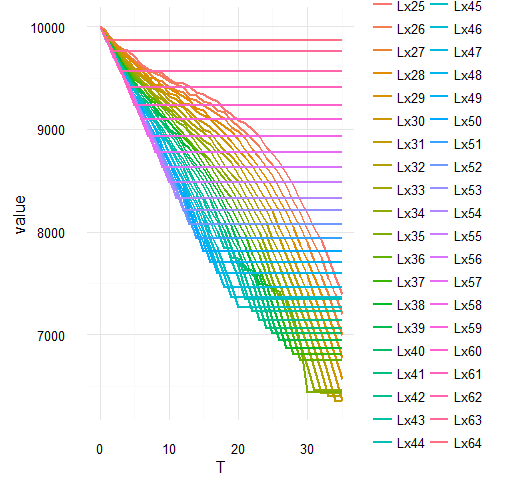
### Nuage de points

Dans un premier temps je réalise une projection des différentes variables sous forme de nuage de points en mettant en évidence les points selon la variable cible notée ici Lx25 (l’évolution d’une population d’âge initial 25ans).

|  |  |
| --- | --- |
|  |  |
|  |  |

Cette première visualisation montre une dynamique de population relativement proche pour les individus de classe d’âge comprise entre 25 et 29 ans et qui devient disparate à mesure que l’on se rapproche de de la cinquantaine. Ce qui pose qui pose à première vue la question de la pertinence de l’utilisation de la table de survie des individus de tranche d’âge supérieure à 50 ans dans le cas pour l’analyse de la dynamique des individus.

### Courbes d’évolution de la dynamique des individus



L’observation de la dynamique de l’évolution pour les différentes classes d’âge considérées suggère une première discrétisation en 4 classes :

* Les individus relativement jeunes d’âge [25 ans ; 35 ans] sur lesquels peuvent être réalisées des analyse sur une période longue d’environ plus de 30ans.
* Les individus d’âge [36,44ans] pour lesquelles la durée d’observation est pertinente sur une période comprise entre 20 et 30 ans.
* Entre [45 ans ,54 ans], la durée d’observation se situe entre 10 et 20ans.
* [55,64ans] durée d’observation va de 1 an à 10 ans. La durée d’observation trop courte pour que l’échantillon soit représentatif.

Les graphes ci-dessus présentent la dynamique évolutive selon cette discrétisation en 4 classes.

|  |  |
| --- | --- |
|  |  |
|  |  |

Réduction de la dimension

En effet sur la base des précédentes analyses nous pouvons considérer que l’observation des individus de classe d’âge comprise entre [25,35] permet d’analyser leur évolution sur une période assez longue pour mieux cerner leurs entrées et sorties d’activités au sein d’une entreprise.

Nous choisissons alors de faire une régression logistique sur la variable cible notée Lx25 offrant la période d’observation la plus longue. Ceci permettra alors d’avoir une idée sur les classes d’âge les plus pertinentes pour calculer les probabilités………………

Cette régression logistique montre les tables issues des observations d’individus d’âge initial de 25 ou de 26 ans sont assez proche.

Dans la suite nous présenterons les résultats issus de l’observation d’individus d’âge 25ans.

# Analyses descriptives

En effet dans notre base de données sont récapitulés pour différents groupes d’individus (10.000) de même âge (10.000), le nombre d’individus restants au sein de l’entreprise sur plusieurs années.

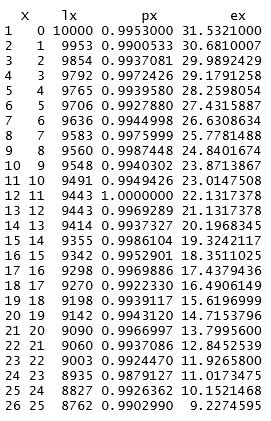
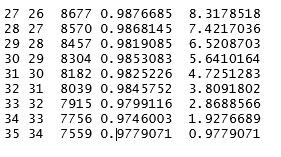
Autrement dit sur toute une période, nous avons les individus qui sont rentrés où sorties de l’entreprise.

Je décide de travailler cette base de données comme une table de mortalité de calculer la probabilité de survie (rester dans l’entreprise), de décès (quitter l’entreprise), espérance de vie active résiduelle.

Pour cela une fonction permettant d’avoir accès à toutes ces informations a été créée et les différents résultats sont présentés pour la table que nous appellerons table de vie active à l’âge de 25 ans notée Lx25.

## Courbe d’évolution

|  |  |
| --- | --- |
|  |  |



Les px représentent les probabilités annuelles qu’un individu d’âge x+25 soit présent dans l’entreprise dans l’année qui suit et les ex représentent les durées de vie résiduelle.

## Estimation du taux de départ

Nous observons que la loi d’évolution pour toutes les observables sont strictement décroissantes.

Une première modélisation simple de cette décroissance peut se faire à travers l’équation différentielle

\_

On note N(t) le nombre d’individus encore présents à l’instant t ; et λ le taux de décroissance. Le but est d’estimer λ(N , t) à partir des données mesurées.