Comment passer des anciennetés aux durées ? Illustration à partir de l'Enquête Famille et Logement.

Vianney Costemalle

Unité des études démographiques et sociales, Insee.

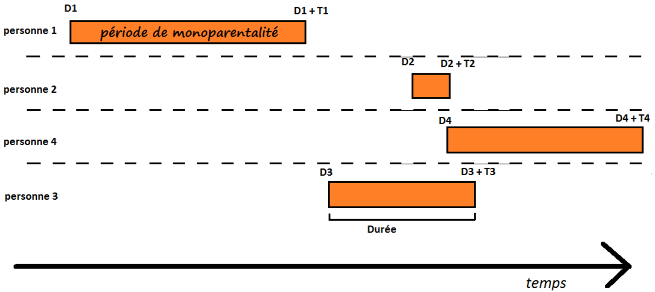
1 avril 2015

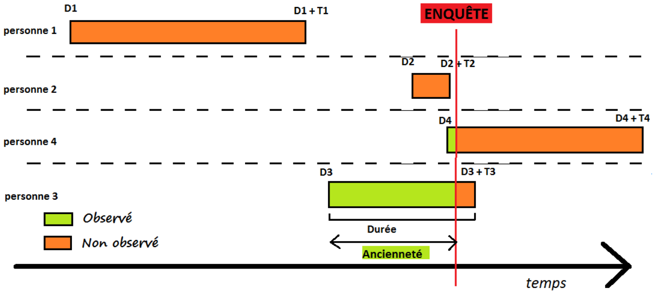
- L'Enquête Famille et Logement (EFL) a été réalisé en 2011 conjointement au recensement de la population, sur 360000 personnes âgées de 18 ans ou plus.
- Elle permet en particulier de connaître le stock de **familles monoparentales** en janvier 2011.

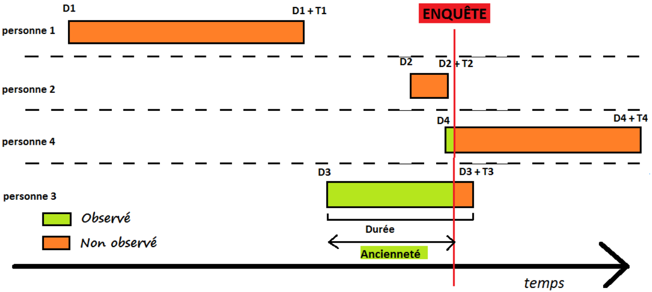
Définition de la famille monoparentale :

Une famille monoparentale est composé d'un parent **sans conjoint cohabitant** ayant des **enfants mineurs** vivant dans le même logement.

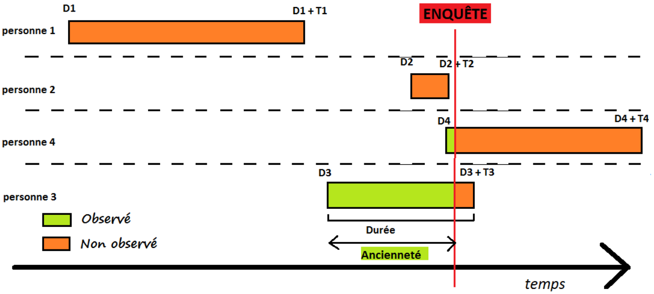
- Grâce à une question sur leur dernier couple, on sait **depuis combien de temps** les adultes concernés sont en famille monoparentale. Mais on ne sais pas **combien de temps ils vont rester** en famille monoparentale.
- Ancienneté : "Depuis combien de temps est-on dans cette situation ?"
- Durée : "Combien de temps dure réellement cette situation ?"


D1
D1 + T1
personne 1
période de monoparentalité
D2
D2 + T2
personne 2
D4
D4 + T4
personne 4
D3
D3 + T3
personne 3
Durée
temps


ENQUÊTE
D1
D1 + T1
personne 1
D2
D2 + T2
personne 2
D4
D4 + T4
personne 4
D3
D3 + T3
personne 3
Observé
Non observé
Durée
Ancienneté
temps



- **Effet de censure** : les anciennetés sont plus courtes que les durées, par définition.



- **Effet de censure** : les anciennetés sont plus courtes que les durées, par définition.
- **Effet de sélection** : les personnes restant longtemps dans la situation ont plus de chance d'être dans cette situation au moment de l'enquête.

Questions :

- Quels sont les liens entre anciennetés et durées ?

Questions :

- Quels sont les liens entre anciennetés et durées ?
- Si les anciennetés du *groupe 1* sont en moyenne plus petites que les ancienntés du *groupe 2*, cela implique-t-il que les durées du *groupe 1* sont aussi en moyenne plus petites que celles du *groupe 2* ?

Questions :

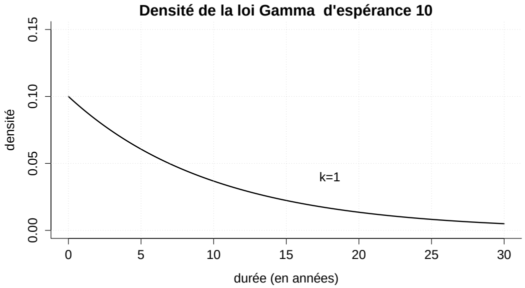
- Quels sont les liens entre anciennetés et durées ?
- Si les anciennetés du *groupe 1* sont en moyenne plus petites que les ancienntés du *groupe 2*, cela implique-t-il que les durées du *groupe 1* sont aussi en moyenne plus petites que celles du *groupe 2* ?
- Peut-on inférer les durées à partir des anciennetés ? Quelles informations supplémentaires sont nécéssaires à cette inférence ?

Notations :

- T : **Durée** de la situation étudiée.
- D : Date de **début** de la situation.
- E : Date de l'**enquête**.
- A : **Ancienneté** de la situation au moment de l'enquête ($A = E - D$).

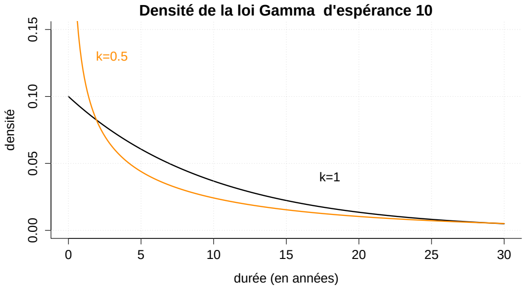
Variable de durée

- On simulera les durées à l'aide de la loi Gamma : $T \sim \text{Gamma}(k, \lambda)$.
- k est le paramètre de *forme* et λ le paramètre d'*échelle*.



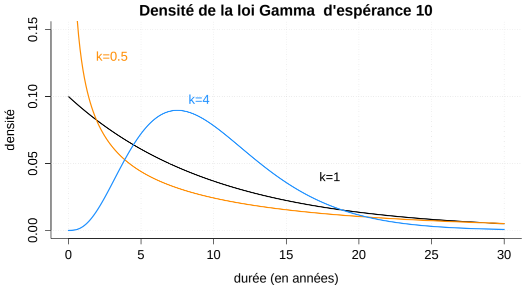
Variable de durée

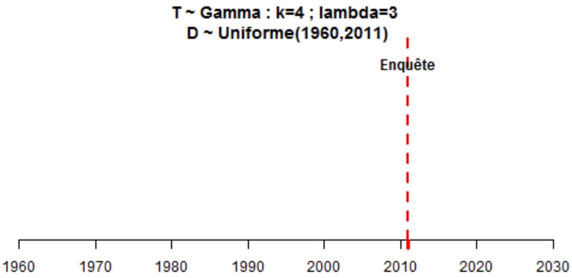
- On simulera les durées à l'aide de la loi Gamma : $T \sim \text{Gamma}(k, \lambda)$.
- k est le paramètre de *forme* et λ le paramètre d'*échelle*.



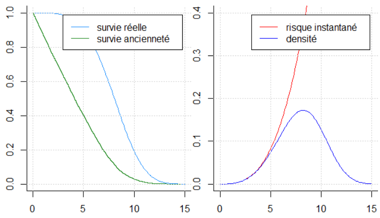
Variable de durée

- On simulera les durées à l'aide de la loi Gamma : $T \sim \text{Gamma}(k, \lambda)$.
- k est le paramètre de *forme* et λ le paramètre d'*échelle*.




T ~ Gamma : k=4 ; lambda=3
D ~ Uniforme(1960,2011)
Enquête
1960
1970
1980
1990
2000
2010
2020
2030

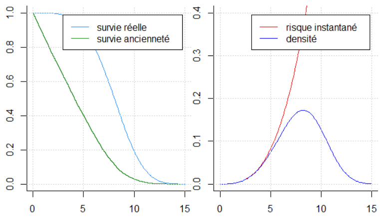
Effet **censure** plus fort ($k > 1$) :



Temps moyen en années	
Ancienneté	4.4
FMP enquête	8.6
Réel	8

7.8% sont en FMP au moment de l'enquête.

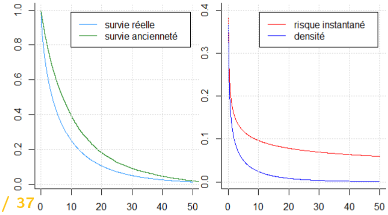
Effet **censure** plus fort ($k > 1$) :



Temps moyen en années	
Ancienneté	4.4
FMP enquête	8.6
Réel	8

7.8% sont en FMP au moment de l'enquête.

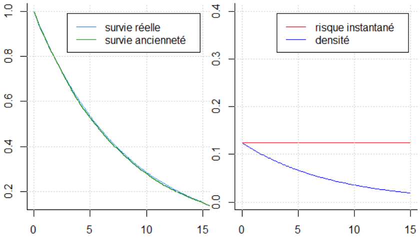
Effet **sélection** plus fort ($k < 1$) :



Temps moyen en années	
Ancienneté	11.6
FMP enquête	23.5
Réel	8

7.7% sont en FMP au moment de l'enquête.

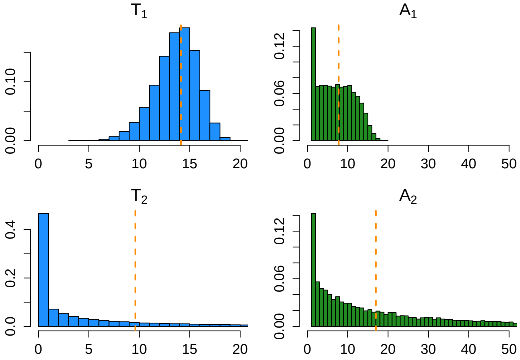
Ces deux effets s'annulent exactement si $k = 1$ ($T \sim$ *exponentielle*) :



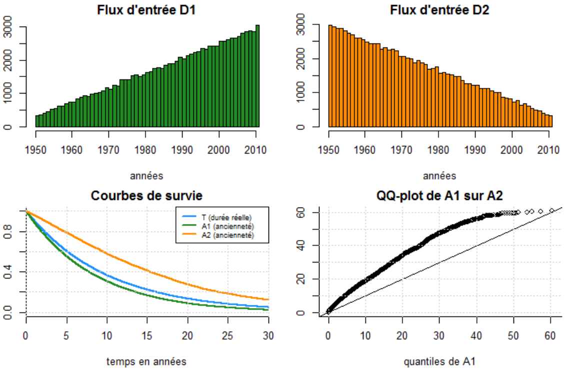
Temps moyen en années	
Ancienneté	7.9
FMP enquête	15.9
Réel	8

7.9% sont en FMP au moment de l'enquête.

- 2 groupes : T_i la durée *groupe i* et A_i son ancienneté.
- On peut avoir $\mathbb{E}[A_1] < \mathbb{E}[A_2]$ MAIS $\mathbb{E}[T_1] > \mathbb{E}[T_1]$.



- Si le deux groupes ont les mêmes durées ($T_1 = T_2$) ...
- ... on peut avoir $\mathbb{E}[A_1] < \mathbb{E}[A_2]$.



L'effet censure et l'effet sélection sont déterminés par :

- La forme de la densité de la variable de durée T.
- Le flux d'entrée en famille monoparentale.

Probabilité d'une observation

- e : date de l'enquête (janvier 2011).
- Les **observations** sont les dates de début de monoparentalité d_1, ... d_n pour **ceux qui sont en famille monoparentale au moment de l'enquête**.
- Données discrètes : on n'observe que l'année de début de monoparentalité (*ex : 1995, 2007, ...*).
- X : covariable représentant des caractéristiques individuelles (*ex : sexe, catégorie sociale, ...*)

Probabilité d'une observation

On note S_T la **survie** de la variable de durée T :

$$S_T(t,x) = \mathbb{P}(T \geq t | X = x)$$

Probabilité d'une observation

On note S_T la **survie** de la variable de durée T :

$$S_T(t,x) = \mathbb{P}(T \geq t | X = x)$$

$$\mathbb{P}(D_i = d_i | \text{FMP lors de l'enquête et } X = x_i)$$

Probabilité d'une observation

On note S_T la **survie** de la variable de durée T :

$$S_T(t, x) = \mathbb{P}(T \geq t | X = x)$$

$$\begin{aligned} & \mathbb{P}(D_i = d_i | \text{FMP lors de l'enquête et } X = x_i) \\ = \quad & \mathbb{P}(D_i = d_i | D_i \leq e < D_i + T_i \text{ et } X = x_i) \end{aligned}$$

Probabilité d'une observation

On note S_T la **survie** de la variable de durée T :

$$S_T(t,x) = \mathbb{P}(T \geq t | X = x)$$

$$\begin{aligned} & \mathbb{P}(D_i = d_i | \text{FMP lors de l'enquête et } X = x_i) \\ = \ & \mathbb{P}(D_i = d_i | D_i \leq e < D_i + T_i \text{ et } X = x_i) \\ = \ & \frac{S_T(e - d_i, x_i)\mathbb{P}(D_i = d_i | X = x_i)}{\int_u S_T(e-u, x_i)\mathbb{P}(D = u | X = x_i)du} \end{aligned}$$

Probabilité d'une observation

On note S_T la **survie** de la variable de durée T :

$$S_T(t,x) = \mathbb{P}(T \geq t | X = x)$$

$$\begin{aligned} & \mathbb{P}(D_i = d_i | \text{FMP lors de l'enquête et } X = x_i) \\ = \; & \mathbb{P}(D_i = d_i | D_i \leq e < D_i + T_i \text{ et } X = x_i) \\ = \; & \frac{S_T(e - d_i, x_i)\mathbb{P}(D_i = d_i | X = x_i)}{\int_u S_T(e - u, x_i)\mathbb{P}(D = u | X = x_i) du} \end{aligned}$$

Hypothèse 1 :

T et D sont indépendantes

Si D est uniforme

Si $D \sim$ *uniforme*, alors la densité de l'ancienneté peut s'écrire :

$$f_A(x) = \frac{S_T(x)}{\mathbb{E}[T]}$$

- si $T \sim$ *exponentielle* , $f_A(x) = f_T(x) \rightarrow$ A et T ont la même densité !

Si D est uniforme

Si $D \sim$ *uniforme*, alors la densité de l'ancienneté peut s'écrire :

$$f_A(x) = \frac{S_T(x)}{\mathbb{E}[T]}$$

- si $T \sim$ *exponentielle* , $f_A(x) = f_T(x) \rightarrow$ A et T ont la même densité !

Modélisation à partir du risque instantané

On note h le **risque instantané** associé à la variable T :

$$h(t,x) = \mathbb{P}(T = t | T \geq t \text{ et } X = x)$$

Modélisation à partir du risque instantané

On note h le **risque instantané** associé à la variable T :

$$h(t,x) = \mathbb{P}(T = t | T \geq t \text{ et } X = x)$$

On a alors :

$$S_T(t,x) = \prod_{\tau=0}^{t-1} (1 - h(\tau, x))$$

Modélisation à partir du risque instantané

On note h le **risque instantané** associé à la variable T :

$$h(t,x) = \mathbb{P}(T = t | T \geq t \text{ et } X = x)$$

On a alors :

$$S_T(t,x) = \prod_{\tau=0}^{t-1} (1 - h(\tau, x))$$

Approche de S.Nickell dans *"Estimating the probability of leaving unemployment" (1979)* :
$h(t,x) = \frac{1}{1+exp(\beta x+\alpha_1 t+\alpha_2 t^2)}$

Modélisation à partir du risque instantané

On note h le **risque instantané** associé à la variable T :

$$h(t,x) = \mathbb{P}(T = t | T \geq t \text{ et } X = x)$$

On a alors :

$$S_T(t,x) = \prod_{\tau=0}^{t-1} (1 - h(\tau,x))$$

Approche de S.Nickell dans *"Estimating the probability of leaving unemployment" (1979)* :
$h(t,x) = \frac{1}{1+exp(\beta x + \alpha_1 t + \alpha_2 t^2)}$ → ça ne convient pas ici. Ce n'est pas assez flexible .

Modélisation à partir du risque instantané

On note h le **risque instantané** associé à la variable T :

$$h(t,x) = \mathbb{P}(T = t | T \geq t \text{ et } X = x)$$

On a alors :

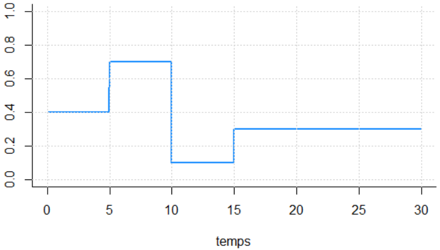
$$S_T(t,x) = \prod_{\tau=0}^{t-1} (1 - h(\tau, x))$$

Approche de S.Nickell dans *"Estimating the probability of leaving unemployment" (1979)* :
$h(t,x) = \frac{1}{1+exp(\beta x + \alpha_1 t + \alpha_2 t^2)}$ → ça ne convient pas ici. Ce n'est pas assez flexible .

Hypothèse 2 : risques instantanés proportionnels

$$h(t,x) = h_0(t) e^{\beta x}$$

On modélise le risque instantané de base, h_0, par une fonction **constante par morceaux**.



On peut choisir :

- la taille du pas élémentaire (5 ans sur la figure).
- le nombre de pas élémentaire (4 pas sur la figure).

Estimation de $\mathbb{P}(D)$.

La probabilité $\mathbb{P}(D)$ est proportionnelle aux **flux d'entrée** en famille monoparentale. $\rightarrow$ on utilise ici une source annexe : l'Étude des Relations Familiales et Intergénérationnelles (ERFI).

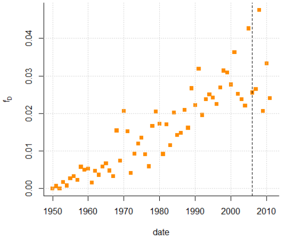


Figure: Estimation de la probabilité d'entrer en famille monoparentale. Champ : personnes âgées de 18 à 73 ans en 2005, France métropolitaine. Source : ERFI vagues 1-3.

Estimation de $\mathbb{P}(D)$.

La probabilité $\mathbb{P}(D)$ est proportionnelle aux **flux d'entrée** en famille monoparentale. $\rightarrow$ on utilise ici une source annexe : l'Étude des Relations Familiales et Intergénérationnelles (ERFI).

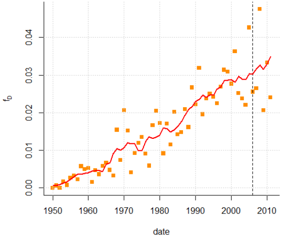


Figure: Estimation de la probabilité d'entrer en famille monoparentale. Champ : personnes âgées de 18 à 73 ans en 2005, France métropolitaine. Source : ERFI vagues 1-3.

Vraisemblance

Vraisemblance des observations $d_1, ..., d_n$

$$\prod_{i=1}^{n} \left[\frac{\hat{f}_D^{Erfi}(d_i, x_i) \prod_{\tau=0}^{2010-d_i} (1 - h_0(\tau) e^{\beta x_i})}{\sum_{u \leq 2010} \left[\hat{f}_D^{Erfi}(u, x_i) \prod_{\tau=0}^{2010-u} (1 - h_0(\tau) e^{\beta x_i}) \right]} \right]$$

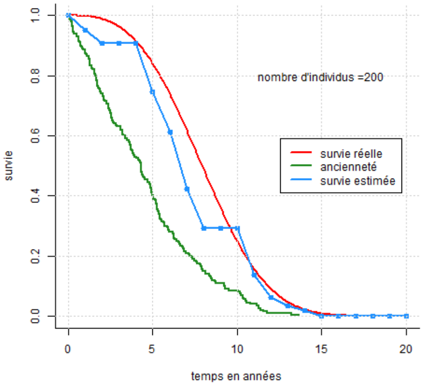


Figure: Effet censure.

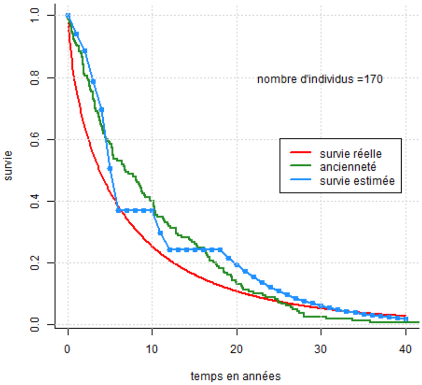
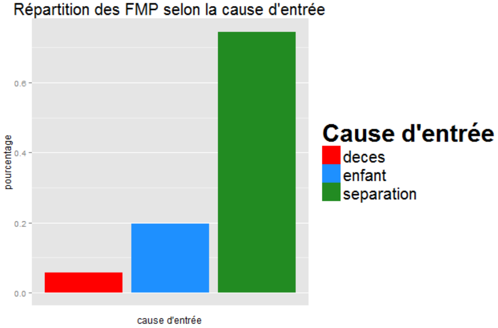


Figure: Effet sélection.

- Sur 360000 adultes interrogés, 12519 sont déterminés comme étant en FMP.
- 1073 hommes et 11446 femmes.

- Sur 360000 adultes interrogés, 12519 sont déterminés comme étant en FMP.
- 1073 hommes et 11446 femmes.

Champ : personnes de 18 ou plus en 2011 étant en FMP lors de l'enquête, France métropolitaine.
Source : EFL



Population entière

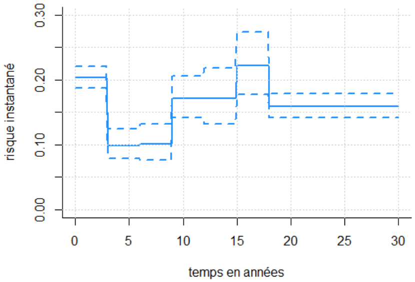


Figure: Estimation du risque instantané de quitter la situation de monoparentalité (source : EFL) .

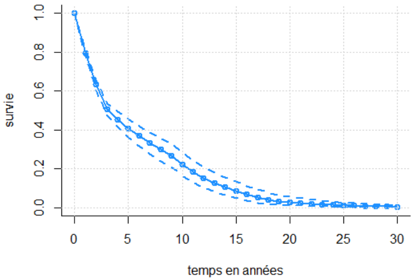


Figure: Estimation de la fonction de survie (source : EFL).

Différence selon le sexe

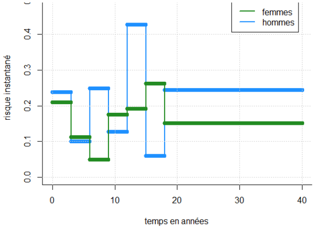


Figure: Estimation du risque instantané de quitter la situation de monoparentalité selon le sexe (source : EFL) .

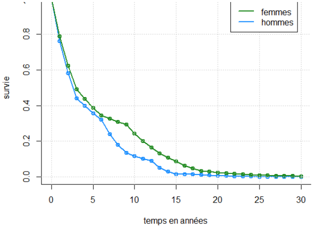


Figure: Estimation de la fonction de survie selon le sexe (source : EFL).

Différence selon le sexe

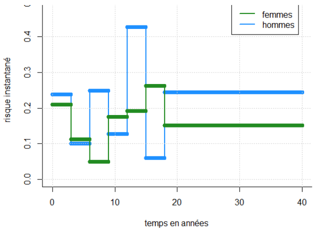


Figure: Estimation du risque instantané de quitter la situation de monoparentalité selon le sexe (source : EFL) .

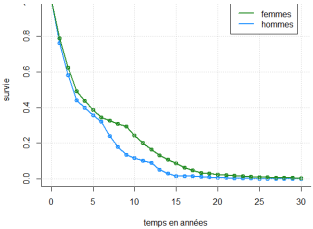


Figure: Estimation de la fonction de survie selon le sexe (source : EFL).

$\rightarrow$ on ne retient pour la suite que l'analyse des femmes en famille monoparentale.

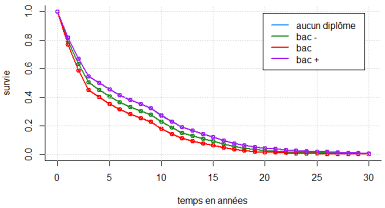
Niveau de diplôme atteint au moment de l'enquête

4 niveaux de diplôme (aucun diplôme, diplômes de niveau inférieur au bac, diplômes équivalents au bac, diplômes de niveau supérieur au bac) .

$$h(t,x) = h_0(t)exp(\beta_{bac-}\mathbb{1}_{(DIP=Bac-)} + \beta_{bac}\mathbb{1}_{(DIP=Bac)} + \beta_{bac+}\mathbb{1}_{(DIP=Bac+)})$$

Coefficient	Valeur	IC à 95%
β_{bac-}	0.12	$(0.061, 0.18)$
β_{bac}	0.25	$(0.18, 0.32)$
β_{bac+}	-0.0005	$(-0.07, 0.07)$

Catégorie de référence : aucun diplôme

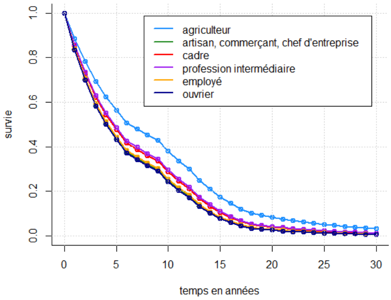


Champ : Femmes de plus de 18 ans en 2011, France métropolitaine. Source : EFL

Catégorie sociale au moment de l'enquête

Coefficient	Valeur	IC à 95%
$\beta_{agriculteur}$	-0.35	$(-0.93, 0.23)$
$\beta_{artisan}$	-0.014	$(-0.15, 0.12)$
β_{cadre}	-0.11	$(-0.23, 0.003)$
$\beta_{prof.inter.}$	-0.14	$(-0.23, -0.05)$
$\beta_{employé}$	-0.03	$(-0.11, 0.045)$

Catégorie de référence : ouvrier

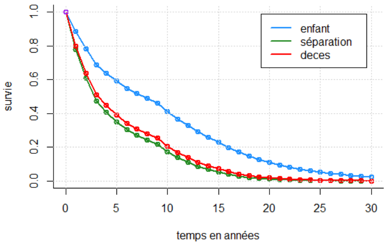


Champ : Femmes de plus de 18 ans en 2011, France métropolitaine. Source : EFL

Cause d'entrée en famille monoparentale

Coefficient	Valeur	IC à 95%
$\beta_{séparation}$	0.64	(0.57, 0.7)
$\beta_{décès}$	0.54	(0.44, 0.63)

Catégorie de référence : enfant



Champ : Femmes de plus de 18 ans en 2011, France métropolitaine. Source : EFL

Comparaison avec ERFI

- avantage : pas d'effet de sélection → modèles de durée classiques avec censure à droite.
- inconvénient : effectif plus restreint (1677 personnes ayant vécu une situation de monoparentalité)

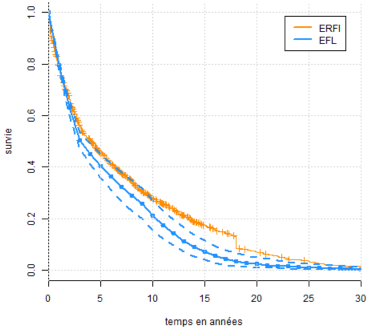


Figure: Comparaison entre la survie estimée avec l'EFL et la survie estimée avec ERFI

Discussion des hypothèses

1. T et D indépendants : les simulations montrent que
   - $corr(T, D) < 0$, on sous-estime la survie
   - $corr(T, D) > 0$, on sur-estime la survie

   On observe dans ERFI une **corrélation négative** entre T et D.

Insee
Mesurer pour comprendre

Discussion des hypothèses

1. T et D indépendants : les simulations montrent que
   - $corr(T, D) < 0$, on sous-estime la survie
   - $corr(T, D) > 0$, on sur-estime la survie

   On observe dans ERFI une **corrélation négative** entre T et D.
2. Risques instantanés proportionnels :

Discussion des hypothèses

1. T et D indépendants : les simulations montrent que
   - $corr(T, D) < 0$, on sous-estime la survie
   - $corr(T, D) > 0$, on sur-estime la survie

   On observe dans ERFI une **corrélation négative** entre T et D.
2. Risques instantanés proportionnels :
3. Covariables indépendantes du temps : la catégorie sociale, le diplôme peuvent changer au cours du temps.

Discussion des hypothèses

1. T et D indépendants : les simulations montrent que
   - $corr(T, D) < 0$, on sous-estime la survie
   - $corr(T, D) > 0$, on sur-estime la survie

   On observe dans ERFI une **corrélation négative** entre T et D.
2. Risques instantanés proportionnels :
3. Covariables indépendantes du temps : la catégorie sociale, le diplôme peuvent changer au cours du temps.
4. Possibilité d'estimer f_D : il faut une source annexe.

Questions ?