PROJET : Durée de chômage

Sommaire

Introduction :

Le but de ce projet est de proposer un modèle de durée sur la sortie de chômage et de tenter d'expliquer la durée passée en situation de chômage des individus en fonction de différentes caractéristiques, à savoir : leur âge, leur situation familiale, leur salaire, leur sexe, s'ils possèdent ou non une assurance chômage, etc. Cette question vaut la peine d'être posée notamment pour aider à tarifer les assurances chômage ou bien à déterminer le montant des aides publiques du chômage. Nous disposons de 2 bases de données de 3343 individus. L'étude a duré 28 mois : au début de l'observation tous les individus sont en situation de chômage et ont une probabilité de sortir du chômage > 0. Nous utiliserons SAS pour analyser ces données en testant différents modèles statistiques. Le but étant, par exemple, de comprendre que certaines variables explicatives sont peu utiles à la résolution du problème dans le modèle final.

Dans la base de données *Base_unemployment_survival,* nous avons la variable *Id*, la variable *event* et la variable *spell* et dans la base de données *ID_databse* nous retrouvons de nouveau la variable *Id* en plus de 7 variables explicatives.

Nous avons donc créé une base de données unique qui rassemble l'ensemble des données individuelles contenues dans les 2 bases de données. Avant de fusionner nos 2 bases, il nous a d'abord fallu les trier par la variable *Id*. Comme les 2 bases contiennent les mêmes individus, nous n'avons pas eu de souci particulier concernant la fusion des 2 bases en une unique base de données.

Notre base de données comporte 3343 individus et 7 variables explicatives :

- *Ui :* indique si l'individu possède une assurance chômage ; Ui=1 si oui et Ui=0 sinon.
- *Logwage :* salaire en logarithme de l'individu
- *Houshead :* indique si l'individu est chef de famille ; houshead=1 si oui et houshead=0 sinon.
- *Married :* indique si l'individu est marié ; married=1 si oui et married=0 sinon.
- *Female :* indique si l'individu est une femme ; female=1 si oui et female=0 sinon.
- *Child :* indique si l'individu a des enfants ; child=1 si oui et child=0 sinon.
- *Age :* âge de l'individu

Les 3 autres variables de la base de données sont :

- *Id* : Variable permettant d'identifier nos individus
- *event* : Variable indiquant si l'évènement « sortie du chômage » est observé ; *event*=1 si l'individu a retrouvé du travail et *event=0* sinon.
- *spell :* Variable de durée (en mois).

Elaboration des données statistiques

Nous élaborerons dans cette partie un ensemble de calculs statistiques nous permettant de décrire les données dont nous disposons. Nous tenterons de décrire la population observée : observe-t-on une population plutôt mariée ? Une population avec des enfants ? Une population âgée ? etc. Nous chercherons également à estimer le temps que met un individu au chômage à retrouver du travail.

Analyse de notre population

Cadre : On observe 3343 individus sur une période de 28 mois. Au début de l'observation, les 3343 individus observés sont au chômage. La variable à expliquer est donc *event* et nous avons 8 variables explicatives : *spell, ui, logwage, houshead, married, female, child et age*. La variable *Id* est notre variable « identifiant » qui permet de distinguer chaque individu. La variable *spell* est une variable de durée, mais pour cette partie elle peut être considérée comme une variable explicative de la variable *event*.

Procédure FREQ

sortie_chomage				
event	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	2270	67.90	2270	67.90
1	1073	32.10	3343	100.00

Proc freq pour la variable event

Cette première table de fréquence sur la variable à expliquer *event* nous permet de constater le nombre de données censurées contenues notre base de données.

Evènements observés : 1073/3343 (données complètes).
Evènements inobservés : 2270/3343 (données incomplètes).

32.10% des individus observés sont sortis du chômage pendant la période d'observation. Pour les 67.90% autres individus, nous n'avons pas pu observer la réalisation de l'évènement « sortir du chômage » sur la période d'observation. On ignore si ces individus ont retrouvé du travail ou non, ce sont des données incomplètes : les individus ont pu retrouver du travail après la période d'observation (données tronquées) ou bien ils ont pu arrêter de répondre à l'étude (données censurées).

Procédure MEANS

Variable d'analyse : spell variable de durée: temps passé au chomage		
Minimum	Moyenne	Maximum
1.0000000	6.2479809	28.0000000

Procédure MEANS

Variable d'analyse : spell variable de durée: temps passé au chomage		
Minimum	Moyenne	Maximum
1.0000000	6.8283319	27.0000000

Figure 1 proc means spell sur l'ensemble de l'échantillon (à gauche) et proc means spell sur les event=0 (à droite)

Sur le tableau de gauche, on peut voir que le temps passé au chômage maximum est de 28 mois, on devine donc que la période d'observation est de 28 mois. Or sur le tableau de droite, concernant les 2270 individus avec *event=0,* la durée de chômage maximum est de 27 mois. Ces 2270 évènements inobservés sont donc tous des censures. Les évènements inobservés signifient donc que l'on a perdu les individus lors de la période d'observation. On ignore s'ils sont encore au chômage à la fin de l'observation. On connaît par contre la durée minimum qu'ils ont passée en situation de chômage.

Variables continues

Nous avons créé des classes pour les variables continues *logwage*, *age* et *spell* afin de simplifier la lecture de données.

classe_logwage	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
[2.7;4]	7	0.21	7	0.21
]4;5]	285	8.53	292	8.73
]5;6]	2156	64.49	2448	73.23
]6;7.7]	895	26.77	3343	100.00

classe_age	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
[20-30]	1319	39.46	1319	39.46
]30-40]	1000	29.91	2319	69.37
]40-50]	649	19.41	2968	88.78
]50;61]	375	11.22	3343	100.00

classe_spell	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
[01;06]	2147	64.22	2147	64.22
]06;12]	687	20.55	2834	84.77
]12;18]	360	10.77	3194	95.54
]18;24]	91	2.72	3285	98.27
]24;28]	58	1.73	3343	100.00

Figure 2 Proc freq variables continues

Libellé	Minimum	Maximum	Moyenne	Mode	Quartile inférieur	Médiane	Quartile supérieur	Ecart-type	Variance
spell	1.0000000	28.0000000	6.2479809	1.0000000	2.0000000	5.0000000	9.0000000	5.6112710	31.4863626
logwage	2.7080500	7.6004000	5.6929944	5.4680600	5.2983200	5.6767500	6.0520900	0.5356591	0.2869307
age	20.0000000	61.0000000	35.4433144	27.0000000	27.0000000	34.0000000	43.0000000	10.6402042	113.2139462

Figure 3 Proc means pour les variables continues

Logwage

Le log salaire de notre échantillon est compris entre 2,7081 et 7,6004. On constate que 64.49% de l'échantillon a un log-salaire compris dans l'intervalle]5 ;6] et en moyenne le log-salaire est de **5.693.** Le rapport de l'écart type est faible (0,536) par rapport à la moyenne (5,693), ce qui signifie que les valeurs ne sont donc pas très dispersées autour de la moyenne.

Age

L'âge de la population est compris entre 20 et 61 ans. La majorité de l'échantillon est plutôt jeune puisque 39.46% de 20-30 ans. En effet, on remarque que l'âge le plus présent dans notre population est 27 ans (mode = 27). L'âge moyen de notre population est en revanche de 35,44 ans**.** L'écart type est de 10,64 pour une moyenne de 35,44 ; la population n'est pas trop dispersée autour de la moyenne. En effet, l'écart interquartile est de 16 pour une étendue de 41.

Spell

Il est plus difficile d'interpréter la variable *spell* puisque plus de deux tiers de notre population sont des données incomplètes pour cette variable. Nous **ne pouvons pas** dire que 64.22% de notre échantillon a passé entre 1 et 6 mois au chômage puisque ces données contiennent aussi des individus censurés qui ont en fait passé « au moins 1 à 6 mois de chômage ». De même, on aimerait dire qu'en moyenne le temps passé au chômage de notre échantillon est de 6.25 mois `(min 1 max 28)`, mais n'oublions pas que notre échantillon contient des données incomplètes, **cette information est donc biaisée.**

Pour tenter d'interpréter ces données, nous avons séparé les individus dont l'information est complète des individus dont l'information est incomplète en créant 2 tables :

- La table *event1* contenant les individus dont on a observé la sortie du chômage (*event=1*)
- La table *event0* contenant les individus dont on n'a pas pu observer la sortie de chômage (*event=0*)

sortie_chomage

event	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
1	1073	100.00	1073	100.00

classe_spell	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
[01;06]	783	72.97	783	72.97
]06;12]	169	15.75	952	88.72
]12;18]	99	9.23	1051	97.95
]18;24]	15	1.40	1066	99.35
]24;28]	7	0.65	1073	100.00

sortie_chomage

event	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	2270	100.00	2270	100.00

classe_spell	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
[01;06]	1364	60.09	1364	60.09
]06;12]	518	22.82	1882	82.91
]12;18]	261	11.50	2143	94.41
]18;24]	76	3.35	2219	97.75
]24;28]	51	2.25	2270	100.00

Figure 4 proc freq spell table event=1 (à gauche) et table event=0 (à droite)

Parmi les 32.10% individus sortis du chômage sur la période d'observation, **exactement** 72.97% ont passé entre 1 à 6 mois au chômage.

Parmi les 67.90% individus censurés, **au moins** 39.92% (22.82% + 11.50% + 3.35% + 2.25%) ont passé plus de 6 mois au chômage.

temps passé au chomage				
spell	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
1	294	27.40	294	27.40
2	178	16.59	472	43.99
3	119	11.09	591	55.08
4	56	5.22	647	60.30
5	104	9.69	751	69.99

Extrait de la proc freq pour la variable spell de la table event=1

On remarque la majorité de la table event1 est sortie du chômage en 1 mois.
Plus exactement, pour la table event1 (données complètes) :
27.40% des individus ont passé 1 mois au chômage
16.59% des individus ont passé 2 mois au chômage
11.09% des individus ont passé 3 ans au chômage
9.69% des individus ont passé 5 ans au chômage

On se demande alors si ce sont les jeunes qui sortent plus rapidement du chômage.

Distribution de la durée de chômage en fonction de l'âge pour la table event1

Fréquence Pourcentage Pctage en ligne Pctage en col.	Table de classe_spell par classe_age				
	classe_age				
classe_spell	[20-30]	]30-40]	]40-50]	]50;61]	Total
[01;06]	321 29.92 41.00 76.79	242 22.55 30.91 71.18	154 14.35 19.67 71.96	66 6.15 8.43 65.35	783 72.97
]06;12]	60 5.59 35.50 14.35	56 5.22 33.14 16.47	36 3.36 21.30 16.82	17 1.58 10.06 16.83	169 15.75
]12;18]	32 2.98 32.32 7.66	34 3.17 34.34 10.00	19 1.77 19.19 8.88	14 1.30 14.14 13.86	99 9.23
]18;24]	4 0.37 26.67 0.96	6 0.56 40.00 1.76	2 0.19 13.33 0.93	3 0.28 20.00 2.97	15 1.40
]24;28]	1 0.09 14.29 0.24	2 0.19 28.57 0.59	3 0.28 42.86 1.40	1 0.09 14.29 0.99	7 0.65
Total	418 38.96	340 31.69	214 19.94	101 9.41	1073 100.00

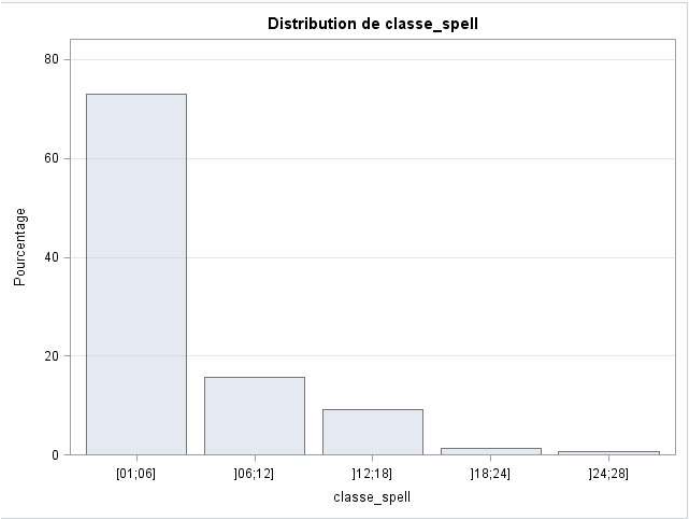
Proc freq spell avec âge pour event1

Parmi les individus qui sont sortis du chômage, on peut voir que ce sont les jeunes qui restent le moins longtemps au chômage : 76.79% des 20-30 ans des *event1* ne restent qu'entre 1 et 6 mois au chômage et 41% des personnes qui ne restent qu'entre 1 et 6 mois au chômage sont des 20-30 ans.

classe_spell	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
[01;06]	783	72.97	783	72.97
]06;12]	169	15.75	952	88.72
]12;18]	99	9.23	1051	97.95
]18;24]	15	1.40	1066	99.35
]24;28]	7	0.65	1073	100.00

Figure 5 proc freq spell pour event=1

Dans notre échantillon, parmi les gens qui sont sortis du chômage pendant l'étude, **72.97% sont sortis durant les 6 premiers mois de chômage.** On remarque que plus la durée de chômage est élevée, moins on observe de personnes qui sortent du chômage.



On voit ci-contre que pour les 1073 personnes avec *event=1* ; la fonction de répartition est décroissante par rapport à la durée passée au chômage.

Figure 7 Distribution des individus dont on a observé la sortie du chômage

En se basant sur cette distribution, on peut dire que plus on reste longtemps au chômage, moins on a de chance de retrouver du travail.

Variables discrètes

assurance_chomage				
ui	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	1495	44.72	1495	44.72
1	1848	55.28	3343	100.00

houshead				
houshead	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	1297	38.80	1297	38.80
1	2046	61.20	3343	100.00

child				
child	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	1838	54.98	1838	54.98
1	1505	45.02	3343	100.00

married				
married	Fréquence	Pourcentage	Fréquence cumulée	Pctage cumulé
0	1384	41.40	1384	41.40
1	1959	58.60	3343	100.00

Proc freq variables discrètes qualitatives

On observe que :

- 55.28% des individus observés, donc la majorité absolue, possède une assurance chômage
- Plus de la moitié de l'échantillon n'a pas d'enfant (54.98%)
- Plus de la moitié de l'échantillon est mariée (58.60%)
- Plus de la moitié de la population observée est chef de famille (61.20%)
  IL MANQUE LA VARIABLE HOMME FEMME

Variable	Libellé	Moyenne	Mode	Ecart-type	Variance
ui	assurance_chomage	0.5527969	1.0000000	0.4972791	0.2472865
houshead	houshead	0.6120251	1.0000000	0.4873617	0.2375214
married	married	0.5860006	1.0000000	0.4926221	0.2426765
female	female	0.3478911	0	0.4763725	0.2269308
child	child	0.4501944	0	0.4975876	0.2475935

Proc means pour les variables discrètes qualitatives

En regardant le mode, ou la moyenne, nous pouvons tout de suite voir que :

- La majorité des individus de notre échantillon possède une **assurance chômage**
- La majorité des individus … sont **chef de famille**
- La majorité des individus … sont **mariés**
- La majorité des individus sont des **hommes**
- La majorité n'a **pas d'enfant**

La variance et l'écart type des variables discrètes sont très proches car ce sont des variables qualitative (2 modalités : oui ou non correspondant à 1 et 0 respectivement). Ce que la *proc means* nous apprend pour les variables qualitatives, la *proc freq* nous l'apprenait aussi.

Analyses sur les corrélations entre les variables

Quel est le profil des individus qui sortent du chômage ?

Procédure CORR

1 Avec les variables :	event
8 Variables :	age ui houshead married female child spell logwage

Statistiques simples

Variable	N	Moyenne	Ecart-type	Somme	Minimum	Maximum	Libellé
event	3343	0.32097	0.46692	1073	0	1.00000	sortie_chomage
age	3343	35.44331	10.64020	118487	20.00000	61.00000	age
ui	3343	0.55280	0.49728	1848	0	1.00000	assurance_chomage
houshead	3343	0.61203	0.48736	2046	0	1.00000	houshead
married	3343	0.58600	0.49262	1959	0	1.00000	married
female	3343	0.34789	0.47637	1163	0	1.00000	female
child	3343	0.45019	0.49759	1505	0	1.00000	child
spell	3343	6.24798	5.61127	20887	1.00000	28.00000	temps_chomage
logwage	3343	5.69299	0.53566	19032	2.70805	7.60040	logwage

Coefficients de corrélation de Pearson, N = 3343
Proba > |r| sous H0: Rho=0

	age	ui	houshead	married	female	child	spell	logwage
event	-0.01419	-0.12391	0.07008	0.08094	-0.00711	0.02697	-0.16013	0.11417
sortie_chomage	0.4120	<.0001	<.0001	<.0001	0.6810	0.1190	<.0001	<.0001

Figure 6 : test de corrélation entre la variable event avec les variables explicatives

Comme la p-value est proche de 0 pour les variables explicatives *ui, houshead, married, spell* et *logwage*, on peut conclure de l'existence d'une corrélation entre ces variables avec la variable à expliquer *event*.

On peut donc établir les corrélations suivantes :

- L'assurance chômage a donc une influence sur la sortie du chômage : les gens qui ont une assurance chômage auront plus tendance à rester au chômage que les gens qui n'en ont pas.
- Les gens qui sont chefs de famille ont plus tendance à sortir du chômage que les gens qui ne le sont pas.
- Les gens mariés ont plus tendance à sortir du chômage que les gens qui ne sont pas mariés.
- Plus on est au chômage depuis longtemps, moins on a de chance de sortir du chômage.
- Plus on a un salaire élevé, plus on a tendance à sortir du chômage.

Quel est le profil des gens qui ne restent pas longtemps au chômage ?

Coefficients de corrélation de Pearson, N = 3343
Proba > |r| sous H0: Rho=0

	age	ui	houshead	married	female	child	logwage
spell	0.15281	0.34376	-0.00507	0.00219	-0.00060	0.00351	0.03957
temps passé au chomage	<.0001	<.0001	0.7693	0.8994	0.9721	0.8391	0.0221

Coefficients de corrélation de Pearson, N = 1073
Proba > |r| sous H0: Rho=0

	age	ui	houshead	married	female	child	logwage
spell	0.07761	0.39618	-0.01060	-0.04437	0.01310	0.00634	-0.00929
temps passé au chomage	0.0110	<.0001	0.7286	0.1464	0.6683	0.8357	0.7612

1e tableau : test de corrélation entre la variable spell et les autres variables explicatives (2e tableau : même chose pour pour event=1)

La p-value est proche de 0 (si on choisit un seuil de 2%) pour les variables *age* et *ui* ce qui nous permet d'établir l'existence d'une corrélation entre ces variables avec la variable *spell* :

- L'âge influence la durée passée en situation de chômage : plus on est âgés, plus le temps passé au chômage va être long
- Les individus qui ont une assurance chômage vont avoir tendance à rester plus longtemps au chômage que les individus qui n'en ont pas.

Pour la variable *spell*, il est plus prudent de se fier au 2e tableau : il a certes moins d'observations mais les données sont complètes. En effet, si on se fiait uniquement au premier tableau, on aurait pu accepter par erreur une corrélation positive entre *logwage* et *spell* avec une probabilité d'erreur inférieure à 3%, ce qui aurait donné comme interprétation erronée : plus le salaire de l'individu est élevé, plus la durée passée en situation de chômage est élevée.

Est-ce que les chefs de familles sont des gens mariés ? Des gens avec des enfants ?

Coefficients de corrélation de Pearson, N = 3343 Proba > |r| sous H0: Rho=0						
	age	ui	married	female	child	logwage
houshead	0.23790	0.04318	0.09602	-0.42246	0.14178	0.28946
houshead	<.0001	0.0125	<.0001	<.0001	<.0001	<.0001

La p-value est proche de 0 (seuil de 1%) pour les variables *age, married, female, child* et *logwage* ce qui nous permet d'établir l'existence d'une corrélation entre ces variables avec la variable *houshead* . On choisit un individu au hasard dans notre échantillon :

- Plus l'individu choisi est âgé, plus la probabilité qu'il soit un chef de famille est élevée.
- Si l'individu choisi est marié, il y a plus de chance pour qu'il soit un chef de famille que s'il n'était pas marié.
- Si l'individu choisi est une femme, il y a moins de chance pour qu'il soit un chef de famille que s'il était un homme.
- Si l'individu choisi a un enfant, il y a plus de chance pour qu'il soit un chef de famille que s'il n'avait pas d'enfant
- Plus le salaire de l'individu choisi est élevé, plus la probabilité qu'il soit un chef de famille est élevée.

Les chefs de famille de notre échantillon sont donc des hommes plutôt âgés, marié, avec des enfants et qui gagnent un salaire élevé.

Qui sont ceux qui gagnent plus d'argent ? Sont-ils âgés, mariés, avec enfants ?

Coefficients de corrélation de Pearson, N = 3343 Proba > |r| sous H0: Rho=0						
	age	houshead	ui	married	female	child
logwage	0.25686	0.28946	0.16217	0.17827	-0.28096	0.00109
logwage	<.0001	<.0001	<.0001	<.0001	<.0001	0.9499

La p-value est proche de 0 (seuil de 0.01%) pour les variables *age, married, female, houshead* et *logwage* ce qui nous permet d'établir l'existence d'une corrélation entre ces variables avec la variable *logwage*. On choisit un individu au hasard dans notre échantillon :

- Plus l'individu choisi est âgé, plus son salaire sera élevé
- Si l'individu choisi est chef de famille, il y a plus de chance pour qu'il ait aussi un salaire élevé comparé à s'il n'était pas chef de famille
- Si l'individu choisi a une assurance chômage, il y a plus de chance pour qu'il ait un salaire élevé comparé à s'il n'en avait pas
- Si l'individu choisi est marié, il y a plus de chance pour qu'il ait un salaire élevé comparé à s'il ne l'était pas

- Si l'individu choisi est une femme, il y a plus de chance pour qu'elle ait un salaire plus faible que si l'individu choisi était un homme

D'après notre échantillon, pour avoir un salaire élevé il faudrait être un homme âgé, marié, chef de famille, avec enfant.

Elaboration de modèles de durée

Modèle non paramétrique : Estimateur de KM

Avant d'utiliser l'estimateur de Kaplan Meyer, il faut veiller à respecter deux hypothèses : la censure doit être non informative et la population étudiée doit être homogène. Si l'une des 2 hypothèses n'est pas respectée, l'estimateur est alors **biaisé**. L'estimateur KM prend en compte les données censurées à droite (c'est-à-dire que la date de sortie du chômage est >= à 28 mois).

On utilise la PROCLIFETEST

Les individus dont on n'a pas pu observer l'évènement « sortie du chômage » peuvent contenir de la censure informative. En effet, on peut penser que les individus qui sont restés au chômage en ont eu marre de répondre au questionnaire alors que ceux qui sortent du chômage vont vouloir le faire savoir. Comme on ne veut pas de censure informative, on va faire notre étude sur la sous-population des individus dont on a observé l'évènement « sortie du chômage ».

Méthode ACT donne la fonction de hasard.

Tests d'homogénéité sur la sous-population event=1 :

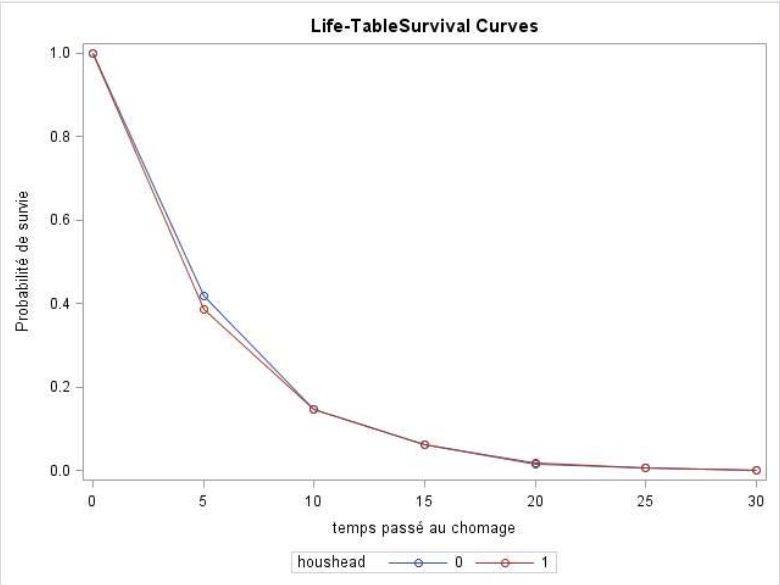
Sous l'hypothèse nulle H0, les strates sont égales (ie : les strates ont des distributions équivalentes). Une P-value proche de zéro s'interprète de cette manière : je n'ai aucune chance de me tromper en rejetant H0, donc je vais rejeter H0. Rejeter H0 revient à accepter que mes populations de strates sont différentes.

Houshead :

Test d'égalité sur niveaux de discrétisation			
Test	Khi-2	DDL	Pr > khi-2
Log-rang	0.1503	1	0.6983
Wilcoxon	0.9472	1	0.3304
-2Log(LR)	0.1175	1	0.7318

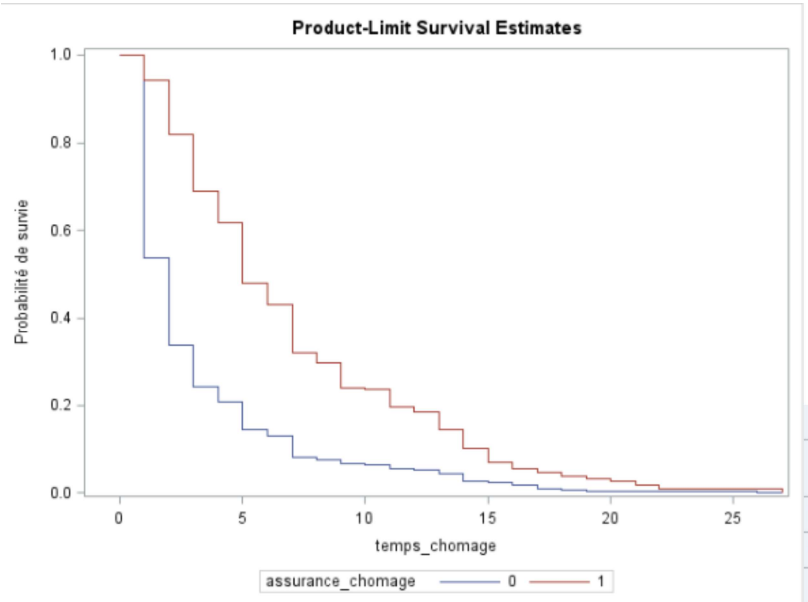
La p-value du test du log-rank et celle du test de wilcoxon sont trop élevées pour pouvoir rejeter l'hypothèse nulle. On ne peut pas conclure que les populations sont différentes, on accepte donc

l'hypothèse nulle : la population est homogène. En effet, quand on regarde les courbes de survie des strates, on voit qu'elles sont très proches.



Assurance chômage :

Les populations de strates sont différentes, en effet on peut voir sur la figure ci-dessous que les tests de Log-rank et de Wilcoxon rejettent l'hypothèse nulle. De plus, on voit très clairement qu'il y a deux courbes de survie distinctes : les gens qui n'ont pas d'assurance chômage (courbe bleue) retrouvent plus vite du travail. En effet, on remarque que la courbe bleue est toujours inférieure à la courbe rouge.



Test d'égalité sur niveaux de discrétisation			
Test	Khi-2	DDL	Pr > khi-2
Log-rang	191.7212	1	<.0001
Wilcoxon	291.3658	1	<.0001
-2Log(LR)	171.5520	1	<.0001

Married :

Logwage :

**** recréer un modèle stratifié avec les nouvelles variables****

Sur la population entière :

Houshead : On remarque que les individus qui sont chef de famille sont retrouvent du travail plus vite et sont plus nombreux à retrouver du travail que les individus qui ne sont pas chef de famille.

Assurance chômage : On remarque que les individus qui ne possèdent pas d'assurance chômage ont tendance à retrouver du travail plus rapidement que les individus qui en possèdent une.

Married : Les individus mariés retrouvent plus rapidement du travail que les individus qui ne sont pas mariés.

Logwage : Les individus qui ont un salaire élevé retrouvent plus vite du travail.

**** recréer un modèle stratifié avec les nouvelles variables****

**** FAIRE UNE CONCLUSION EN COMPARANT LES RESULTATS ENTRE SOUS POP ET POP ENTIERE : le biais est-il significatif ?**

- **existence censure informative**
- **exitence hétérogénéité population ****

Modèle paramétrique

Test de Wald avec p-value faible : je n'ai aucune chance de me tromper en rejetant l'hypothèse nulle « tous les coefficients sont nuls ». Donc en fait j'accepte qu'au moins 1 coefficient est non nul, donc que le modèle a un pouvoir explicatif.

Fonctions de risque : Y = log du temps de défaillance

- Distribution exponentielle : on suppose que le risque instantané (probabilité qu'il survienne aujourd'hui sachant qu'il n'est pas survenu avant) est une constante
- Distribution de Weibull : gamma < 1 (pour le chômage par exemple), alors fonction de risque instantané décroissante. gamma >1 (pour la mort par exemple), alors la fonction de risque instantané est croissante. si gamma=1, la fonction de risque instantané est constante.
- Distribution log-logistique :

Noncensored values : celles qui sont réellement parties
Right censored values : on n'a pas vu s'ils sont partis car la période d'observation est trop faible
Left censored values : ils n'étaient pas là au début de l'analyse, ils rentrent en cours.
Number of observations read : le nombre d'individus total de notre analyse
Missing values : les observations censurées d'une façon ou d'une autre

Une statistique négative (pour modèle weibull, exp, logistic) => plus la variable est élevée, moins l'évènement se produit tôt (donc plus il se produit tard)

Choix de la distribution : Nous n'utiliserons pas la distribution exponentielle puisque notre fonction de risque instantanée n'est pas une constante, elle est décroissante avec le temps.

Modèle de Cox : semi-paramétrique

On utilise la *Proc PHREG* pour faire modèle de Cox. On dit que le modèle de Cox est un modèle semi-paramétrique car on retrouve du KM et du MV.

Variable Child :

Test Wald : 0.2584. On ne peut donc pas rejeter l'hypothèse nulle : La variable Child n'est pas significative.

Variable Female :

Test Wald : 0.7419. On ne peut donc pas rejeter l'hypothèse nulle : La variable Female n'est pas significative.

Variable Married :

Test de Wald de nullité de tous les coefficients : 0.0001 ; On peut donc rejeter l'hypothèse nulle (valeur inférieure au seuil théorique de 5%): La variable Married est donc significative.
La valeur estimée du coefficient est 0.24477. On peut donc conclure que les gens mariés ont plus tendance à retrouver du travail tôt.

Variable Houshead :

Test de Wald : 0.0004; On peut donc rejeter l'hypothèse nulle, la variable Houshead est significative.
La valeur estimée du coefficient est 0.22712. On peut donc conclure que les gens qui sont « chef de famille » ont plus tendance à retrouver du travail tôt.

Variable ui :

Test de Wald : <.0001
La valeur estimée du coefficient est -0.92296. On peut donc conclure que les gens qui ont une assurance chômage ont plus tendance à retrouver du travail tard.

Variable logwage :

Test de Wald : $<.0001$
La valeur estimée du coefficient est 0.25776. On peut donc conclure que plus on a un salaire élevé, plus on retrouve du travail tôt.

Variable âge :

Test de Wald : $<.0001$
La valeur estimée du coefficient est -0.01144. On peut donc conclure que plus on est jeune, plus on retrouve du travail tôt.

Modèle de Cox adapté à l'analyse de la durée passée en situation de chômage : on garde toutes les variables sauf female et child.