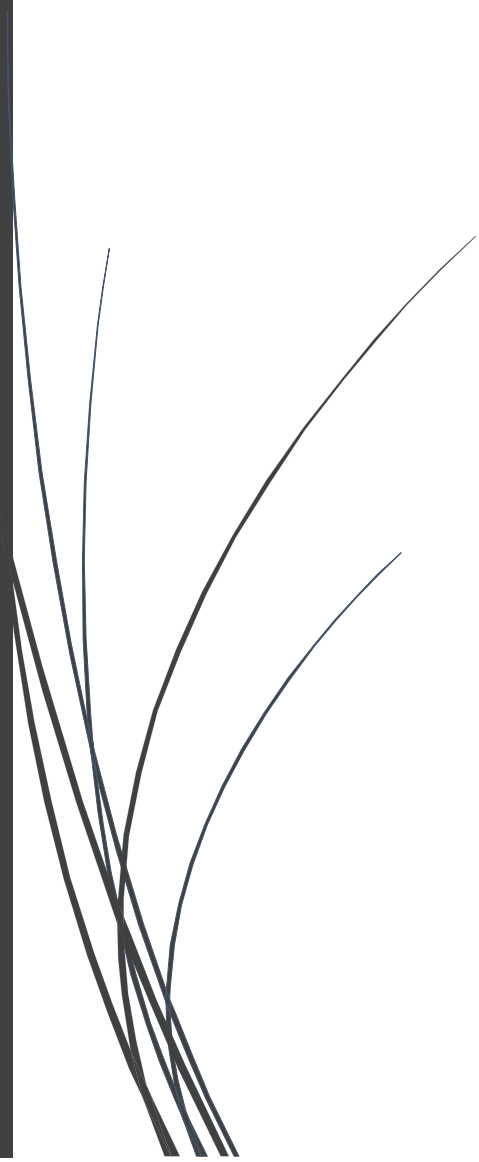
MODELISATION STATISTIQUE

STA110 - DEVOIR N°3

A l'attention de Monsieur Luan JAUPI.



Guillaume Chevron

le cnam

Table des matières

Analyse de la variance à mesures répétées

Un laboratoire souhaite commercialiser un médicament visant à réduire le rythme cardiaque de ses patients.

Pour ce faire, l'équipe de recherche décide de procéder en amont à une phase de vérification des effets sur trois produits : deux médicaments potentiels (A et B) ainsi que d'une solution de contrôle (C). Au cours de cette expérimentation, le rythme cardiaque d'un échantillon de vingt-quatre patients (sujets) est mesuré par intervalle de temps régulier (T0, T2, T4 et T6) après administration de chacun des médicaments.

L'objectif affiché par l'équipe de recherche est de comparer les moyennes des rythmes cardiaques à différents instants de temps régulier, et de comparer la performance de chacun des médicaments observés à partir d'une analyse de la variance à mesures répétées. Elle fait le choix d'un modèle mixte en prenant en compte, en plus des effets fixes (produits et temps), un facteur aléatoire (sujets) :

$$Y = (\text{Intercept}) + (\text{Produit}) + (\text{Temps}) + (\text{Produit*Temps}) + (\text{Sujet})$$

Après avoir construit le plan d'expérience relatif à la phase de vérification en amont, l'équipe de recherche effectue ensuite les mesures souhaitées sur les vingt-quatre sujets observés. Elle obtient ainsi les résultats suivants :

Sujet	Produit	T0	T2	T4	T6
1	A	72	86	81	77
2	B	85	86	83	80
3	C	69	73	72	74
4	A	78	83	88	81
5	B	82	86	80	84
6	C	66	62	67	73
7	A	71	82	81	75
8	B	71	78	70	75
9	C	84	90	88	87
10	A	72	83	83	69
11	B	83	88	79	81
12	C	80	81	77	72
13	A	66	79	77	66
14	B	86	85	76	76
15	C	72	72	69	70
16	A	74	83	84	77
17	B	85	82	83	80
18	C	65	62	65	61
19	A	62	73	78	70
20	B	79	83	80	81
21	C	75	69	69	68
22	A	69	75	76	70
23	B	83	84	78	81
24	C	71	70	65	63

A partir de ces résultats, l'équipe de recherche souhaite déterminer le meilleur modèle de structure de covariance grâce à une comparaison des critères de Schwarz (BIC) et des critères d'Akaike (AIC). Elle procède alors à une analyse statistique sous SAS à partir du code suivant :

```
/*Identification du modèle autorégressif d'ordre 1 avec les meilleurs BIC/AIC*/
proc mixed data=exercice_4c3 ;
title 'Structure de covariance choisie : Autorégressif (1)' ;
class Sujet Prod Temps ; /*Facteurs observés*/
model y = Prod Temps Prod*Temps / s outpred=res ; /*Modèle (partie fixe)*/
random Sujet(Prod) ; /*Facteur aléatoire (avec imbrication)*/
repeated Temps /*Facteur répété (temps)*/ / type=AR(1) /*Structure de covariance choisie*/
subject=Sujet(Prod) group=Prod ;
lsmeans Prod*Temps /slice=Temps diff ; /*Comparaison des moyennes*/
run ;
```

L'équipe de recherche obtient alors les résultats suivants :

Tests d'ajustement	Autorégressif (1)	Toeplitz	Compound Symmetry
-2 log-vraisemblance restreinte	483.7	481.7	488.8
AIC (critère d'Akaike)	489.7	491.7	494.8
AICC (critère d'Akaike corrigé)	490.0	492.4	495.1
BIC (critère de Schwarz)	493.2	497.6	498.3

Les mesures étant régulièrement espacées dans le temps, le modèle autorégressif d'ordre 1 semble être le plus indiqué pour cette expérimentation car les valeurs de son BIC (et de son AIC) sont les plus faibles. A noter également que le modèle de covariance Compound Symmetry ne dépend pas du temps, et n'est donc pas très pertinent dans le cas présent. L'estimation des paramètres de covariance donne les résultats suivants :

- Estimation de la variance du facteur aléatoire Sujet(Prod) : $\sigma^2_B = 20{,}1172$
- Estimation du paramètre ρ du modèle autorégressif d'ordre 1 AR(1) : $\rho = 0{,}5424$
- Estimation de la variance des résidus : $\sigma^2 = 12{,}6439$

L'équipe de recherche effectue ensuite des tests globaux afin de déterminer si les effets fixes observés sont statistiquement significatifs, et obtient les résultats suivants :

Effet	DLL num.	DLL den.	Valeur F	Pr > F
Produit	2	21	6.16	0.0078
Temps	3	63	15.24	<.0001
Produit*Temps	6	63	12.87	<.0001

Elle fait donc les conclusions suivantes :

- Le produit a un effet très significatif : $\text{p-value} = 0{,}0078 < 0{,}01$
- Le temps a un effet hautement significatif : $\text{p-value} < 0{,}001$
- L'interaction entre le produit et le temps a un effet hautement significatif : $\text{p-value} < 0{,}001$

Une fois les tests globaux analysés, l'équipe de recherche procède ensuite à des tests individuels sur ces mêmes effets fixes afin d'obtenir des résultats affinés :

Effet	Produit	Temps	Estimation	Erreur-type	DDL	Valeur du test t	Pr > |t|
Intercept			71.0000	2.0236	21	35.09	<.0001
Prod	A		2.1250	2.8619	21	0.74	0.4660
Prod	B		8.7500	2.8619	21	3.06	0.0060
Prod	C		0	.	.	.	.
Temps		0	1.7500	1.6299	63	1.07	0.2870
Temps		2	1.3750	1.4936	63	0.92	0.3608
Temps		4	0.5000	1.2026	63	0.42	0.6790
Temps		6	0	.	.	.	.
Prod*Temps	A	0	-4.3750	2.3050	63	-1.90	0.0623
Prod*Temps	A	2	6.0000	2.1123	63	2.84	0.0061
Prod*Temps	A	4	7.3750	1.7008	63	4.34	<.0001
Prod*Temps	A	6	0	.	.	.	.
Prod*Temps	B	0	0.2500	2.3050	63	0.11	0.9140
Prod*Temps	B	2	2.8750	2.1123	63	1.36	0.1783
Prod*Temps	B	4	-1.6250	1.7008	63	-0.96	0.3430
Prod*Temps	B	6	0	.	.	.	.
Prod*Temps	C	0	0	.	.	.	.
Prod*Temps	C	2	0	.	.	.	.
Prod*Temps	C	4	0	.	.	.	.
Prod*Temps	C	6	0	.	.	.	.

Les estimations ayant des valeurs nulles correspondent aux conditions de centrage du logiciel SAS, utilisé par l'équipe de recherche dans le cadre de cette expérimentation : dernière estimation nulle pour chacun des effets (Prod et Temps), dernière colonne et dernière ligne nulle pour la matrice d'interaction (Prod*Temps).

L'analyse des effets indique que les résultats suivants :

- Le produit A seul n'est pas significatif (p-value > 0,05)
- Le produit B seul est très significatif (p-value = 0,0060 < 0,01)
- Le temps seul n'est pas significatif, peu importe l'intervalle de mesure observé
- Le produit A est très significatif en fonction du temps (p-value < 0,01 pour T > 0)
- Le produit B n'est pas significatif en fonction du temps (p-value > 0,05 pour tout T)

Le modèle prend donc la représentation symbolique suivante : $Y = I + (\text{Produit A})*(\text{Temps} > 0)$

En comparant les moyennes des rythmes cardiaques à différents instants réguliers, l'équipe de recherche peut donc conclure que le médicament A est véritablement efficace dans le temps, ce qui n'est pas le cas du médicament B.

Méthodes paramétriques et non-paramétriques pour l'étude des durées de vie

La société ACE-Transfo, spécialisé dans la fabrication de transformateurs électriques haute tension, vient récemment de signer une charte de qualité avec sa clientèle, garantissant une durée de vie de l'ensemble de ses composants d'au moins 500 heures. Tout composant n'atteignant pas ce seuil se verra remplacé gratuitement.

Afin d'estimer le nombre de transformateurs qui cesseraient de fonctionner au bout de 500 heures d'activité, et ainsi de pouvoir planifier la production des unités nécessaires pour leurs remplacements, la société ACE-Transfo décide dans un premier temps de mener une analyse de survie à partir de méthodes non-paramétriques, ne faisant donc aucune hypothèse quant à la loi de distribution des défaillances.

Pour ce faire, la société mobilise en interne son équipe technique afin de réaliser une phase de test. Cette dernière est effectuée sur un échantillon de 180 transformateurs parmi lesquels 158 suspensions (censures) et 22 durées de vie (défaillances) sont observées.

L'équipe technique obtient les résultats suivants :

Temps	Censure	Temps	Censure	Temps	Censure	Temps	Censure	Temps	Censure	Temps	Censure
10	0	1982	1	2461	1	2600	1	3388	1	4066	1
314	0	1990	1	2479	1	2649	1	3406	1	4072	1
730	0	2002	1	2479	1	2671	1	3444	1	4073	1
740	0	2012	1	2480	1	2824	1	3462	1	4076	1
990	0	2022	1	2482	1	2882	1	3486	1	4077	1
1046	0	2052	1	2484	1	2902	1	3498	1	4125	1
1570	0	2053	1	2485	1	2915	1	3502	1	4383	1
1870	0	2073	1	2486	1	2935	1	3513	1	4392	1
2020	0	2115	1	2487	1	2937	1	3526	1	4403	1
2040	0	2271	1	2491	1	2945	1	3532	1	4404	1
2096	0	2341	1	2492	1	2949	1	3540	1	4422	1
2110	0	2371	1	2498	1	2950	1	3549	1	4459	1
2177	0	2375	1	2501	1	2956	1	3550	1	4466	1
2306	0	2381	1	2501	1	2964	1	3554	1	4469	1
2690	0	2392	1	2509	1	2968	1	3559	1	4478	1
3200	0	2394	1	2510	1	2971	1	3562	1	4483	1
3360	0	2399	1	2525	1	2978	1	3570	1	4486	1
3444	0	2403	1	2529	1	2979	1	3830	1	4487	1
3508	0	2404	1	2542	1	2980	1	3871	1	4499	1
3770	0	2411	1	2545	1	2989	1	3897	1	4518	1
4042	0	2419	1	2549	1	2995	1	3936	1	4559	1
4186	0	2422	1	2551	1	3017	1	3969	1	4562	1
1796	1	2431	1	2552	1	3018	1	3985	1	4567	1
1831	1	2432	1	2560	1	3034	1	4016	1	4578	1
1840	1	2442	1	2567	1	3035	1	4020	1	4586	1
1890	1	2445	1	2569	1	3042	1	4020	1	4594	1
1927	1	2447	1	2575	1	3069	1	4027	1	4597	1
1944	1	2449	1	2583	1	3106	1	4033	1	4618	1
1956	1	2450	1	2589	1	3142	1	4035	1	4621	1
1982	1	2454	1	2599	1	3146	1	4040	1	4663	1

Ayant fait le choix dans un premier temps de ne pas faire de supposition quant à la loi de distribution de T, l'équipe technique décide d'opter pour la méthode de Kaplan-Meier. Cette méthode peut être appliquée à ces données du fait de la petite taille de l'échantillon choisi et de l'obtention d'observations individuelles. L'équipe technique obtient alors la table de survie suivante[1] :

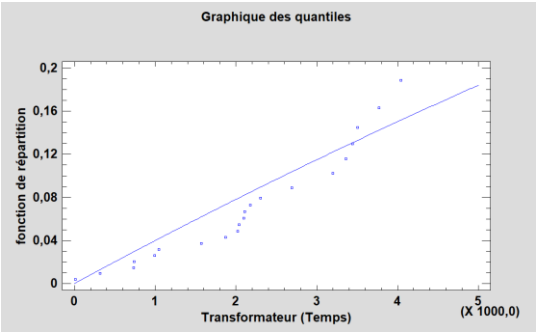
Observation	Temps	Statut	Nombre de présents	Survie cumulée	Erreur-type	Risque cumulé
1	10	DEFAILLANCE	179	0,9944	0,0055	0,0056
2	314	DEFAILLANCE	178	0,9889	0,0078	0,0112
3	730	DEFAILLANCE	177	0,9833	0,0095	0,0168
4	740	DEFAILLANCE	176	0,9778	0,011	0,0225
5	990	DEFAILLANCE	175	0,9722	0,0122	0,0282
6	1046	DEFAILLANCE	174	0,9667	0,0134	0,0339
7	1570	DEFAILLANCE	173	0,9611	0,0144	0,0397
8	1870	DEFAILLANCE	169	0,9555	0,0154	0,0456
9	2020	DEFAILLANCE	159	0,9495	0,0164	0,0518
10	2040	DEFAILLANCE	157	0,9435	0,0174	0,0582
11	2096	DEFAILLANCE	153	0,9374	0,0183	0,0647
12	2110	DEFAILLANCE	152	0,9312	0,0192	0,0713
13	2177	DEFAILLANCE	150	0,9251	0,02	0,0779
14	2306	DEFAILLANCE	148	0,9188	0,0208	0,0846
15	2690	DEFAILLANCE	94	0,9092	0,0228	0,0952
16	3200	DEFAILLANCE	66	0,8956	0,0261	0,1103
17	3360	DEFAILLANCE	65	0,882	0,0291	0,1255
18	3444	DEFAILLANCE	62	0,868	0,0318	0,1415
19	3508	DEFAILLANCE	56	0,8528	0,0347	0,1592
20	3770	DEFAILLANCE	45	0,8343	0,0386	0,1812
21	4042	DEFAILLANCE	31	0,8082	0,0453	0,2129
22	4186	DEFAILLANCE	24	0,7759	0,0538	0,2538

Les estimations obtenues indiquent que seules les deux premières observations n'ont pas atteint la barre des 500 heures, soit près de 1,11% des composants présents au sein de l'échantillon. La troisième observation, se situant au-dessus de ce seuil, enregistre une défaillance au bout de 730 heures avec une probabilité associée de 0,01168 (i.e. avec un degré de confiance de 0,9833). Aussi, la société ACE-Transfo devra donc produire 1,11% transformateurs supplémentaires (i.e. deux au sein de l'échantillon choisi) afin de garantir le remplacement des composants défaillants auprès de sa clientèle.

En s'appuyant sur de la documentation scientifique spécialisée, l'équipe technique décide de partir de l'hypothèse que les défaillances sont distribuées selon une loi de Poisson. Dans la mesure où le taux de défaillance est constant, la loi de survie suit une loi exponentielle.

Les probabilités d'obtenir une valeur inférieure à un seuil préalablement choisi donc les suivantes :

X	Probabilité	Remplacements dans l'échantillon
500,0	0,0201447	4 (3,63)
600,0	0,0241247	5 (4,34)
700,0	0,0280885	6 (5,06)
800,0	0,0320362	6 (5,77)



Pour un temps de fonctionnement fixé à un seuil de 500 heures, et en utilisant une méthode paramétrique, l'équipe technique estime donc qu'il faudra produire près de 2,01% composants supplémentaires, soit quatre transformateurs au sein de l'échantillon, pour garantir le remplacement des pièces défaillantes stipulé par la charte de qualité.

[1] Pour des raisons de lisibilité, les observations censurées ont été retirées de la table de survie.

Comparaison de modèles de régression des durées de vie

Le bureau fédéral des prisons (FBOP), dépendant du Département de la Justice des États-Unis et chargé de l'administration des prisons fédérales américaines, souhaite procéder à une étude statistique sur un échantillon d'individus afin de mieux comprendre le phénomène du retour en prison. L'objectif de cette étude est d'analyser les trois enjeux suivants : la prévalence du retour en prison[2], la vitesse du retour en prison[3] et les facteurs ou caractéristiques auxiliaires qui contribuent statistiquement au retour en prison.

Pour ce faire, une équipe de spécialistes est mobilisée par le FBOP afin d'identifier une dizaine de variables clés à analyser, puis de mener une période d'observation de 52 semaines (une année complète) sur un échantillon de 432 détenus libérés de la prison de l'état de Maryland. Après plusieurs mois de collecte d'informations et d'échanges avec des spécialistes du milieu carcéral, l'équipe identifie finalement dix variables :

1. La durée de temps en semaine jusqu'à la première arrestation après la sortie de prison (Week),
2. L'existence d'une arrestation (Arrest)
3. La présence ou non d'une aide financière (Fin)
4. L'âge en année au moment de la sortie de prison (Age)
5. L'ethnie (Race)
6. L'expérience de travail à temps plein (Wexp)
7. Le statut marital (Mar)
8. L'existence ou non d'une libération sous parole (Paro)
9. Le nombre de précédentes condamnations (Prio)
10. Le niveau d'étude (Edu)

Compte tenu du nombre de variables et de la nature des enjeux à étudier, l'équipe de spécialistes décide de comparer plusieurs les modèles de régression des durées de vie suivants : le modèle de Cox des risques proportionnels, le modèle logistique et le modèle log position – log échelle à partir d'un ajustement de Weibull.

Plusieurs séries de tests sur les rapports de vraisemblance sont donc effectuées pour chacune des méthodes afin d'exclure les facteurs statistiquement non-significatifs, c'est-à-dire dont les valeurs de probabilité sont inférieures à 0,05.

M_1 : Modèle de Cox

Facteur	Khi-carré	Ddl	Proba.
age	13,5967	1	0,0002
prio	10,3482	1	0,0013

M_2 : Modèle logistique

Facteur	Khi-carré	Ddl	Proba.
age	13,6941	1	0,0002
prio	8,41833	1	0,0037

M_3 : Modèle Weibull

Facteur	Khi-carré	Ddl	Proba.
age	13,6222	1	0,0002
prio	10,5283	1	0,0012

L'équipe peut ensuite identifier les estimations des coefficients pour chacun des facteurs retenus :

M_1 : Modèle de Cox

Paramètre	Estimation	Erreur
age	-0,0691492	0,0
prio	0,0944002	0,0

M_2 : Modèle logistique

Paramètre	Estimation	Erreur
constante	0,502177	0,565868
age	-0,077957	0,023027
prio	0,104646	0,0359849

M_3 : Modèle Weibull

Paramètre	Estimation	Erreur
constante	3,84643	0,356555
age	0,049544	0,0154131
prio	-0,06832	0,020063
sigma	0,716565	0,0640341

Ces estimations permettent de déterminer les équations des modèles ajustés :

- Pour le modèle de Cox, la fonction de risque s'écrit comme suit :
  $M_1\ (Cox) : h(t|x) = h(t|0)*\exp(-0{,}0691492*age + 0{,}0944002*prio)$
- Pour le modèle logistique, l'équation du modèle ajusté est la suivante :
  $M_2\ (Log) : arrest = \exp(eta)/(1+\exp(eta))$ où $eta = 0{,}502177 - 0{,}0779574*age + 0{,}104646*prio$
- Enfin, pour le modèle log position – log échelle, l'équation du modèle ajusté est la suivante :
  $M_3\ (W) : week = \exp(3{,}84643 + 0{,}049544*age - 0{,}06832*prio)$

[2] Combien de personnes retournent-elles en prison après avoir été libérées ?
[3] Au bout de combien de temps les personnes retournent-elles en prison après avoir été libérées ?

L'écriture de ces trois modèles ajustés permet à l'équipe de spécialistes d'effectuer plusieurs constats.

Les modèles de Cox et logistique indiquent que :

- Pour chaque année supplémentaire au moment de la sortie de prison, le risque relatif de réincarcération diminue d'environ 7%
- Pour chaque condamnation précédente supplémentaire, le risque relatif de réincarcération augmente d'environ 10%

Le modèle log position – log échelle indique que :

- Pour chaque année supplémentaire au moment de la sortie de prison, la durée de temps en semaine jusqu'à la première arrestation après la sortie de prison augmente d'environ 5%
- Pour chaque condamnation précédente supplémentaire, la durée de temps en semaine jusqu'à la première arrestation après la sortie de prison augmente d'environ 7%

Il est alors possible de procéder à de premières conclusions quant aux enjeux initiaux :

1. Prévalence du retour en prison : 26,39% des individus de l'échantillon ont été de nouveau incarcérés après avoir été libérés
2. Vitesse du retour en prison :
    a. Plus la personne est jeune, plus le retour en prison est rapide
    b. Plus la personne a été condamnée par le passé, plus le retour en prison est rapide
3. Facteurs ou caractéristiques auxiliaires qui contribuent statistiquement au retour en prison :
    a. L'âge au moment de la sortie de prison
    b. Le nombre de précédentes condamnations