Master « Mathématiques Appliquées & Statistique »
Parcours Science des données

Modélisation Statistique

A l'attention de Monsieur Luan JAUPI,
Maître de conférences au Conservatoire National des Arts et Métiers

Guillaume CHEVRON
27/05/2019

le cnam

Table des matières

1. Régression Simple

Un musée, dont un grave incendie a provoqué la fermeture pendant plusieurs mois, doit procéder à des travaux de grande ampleur afin de réparer les installations touchées. Sa fréquentation hebdomadaire ayant été directement impactée par cet incident, les propriétaires du musée souhaitent être indemnisés du préjudice subi par leur compagnie d'assurance. Ils cherchent donc à déterminer une estimation de la fréquentation totale qui aurait été atteinte si l'incendie n'avait pas eu lieu. Cette estimation, effectuée à partir des données de fréquentation mesurées pour un parc d'attraction avoisinant, sera utilisée pour déterminer le niveau d'indemnisation auquel le musée a droit.

Les propriétaires du musée suggèrent alors d'utiliser la période post-incendie, c'est-à-dire les données les plus récentes, car de nouvelles fonctionnalités effectuées pour la réouverture du musée ont amélioré sensiblement les chiffres de fréquentations. A l'inverse, la compagnie d'assurance souhaite prendre en compte la période pré-incendie, c'est-à-dire les données plus anciennes, pour estimer la fréquentation totale perdue des suites de l'incendie. L'objectif de cette étude est donc de comparer les coefficients d'un modèle de régression simple, avant et après l'incendie.

Les données collectées dans le cadre de cette étude correspondent à la fréquentation hebdomadaire du musée (variable dépendante) ainsi que d'un parc d'attraction avoisinant (variable quantitative). Le jeu de donnée initial comprend 205 individus, correspondant aux 205 semaines observées. Cependant, seules les deux périodes durant lesquelles les deux installations fonctionnaient conjointement sont retenues dans le cadre de cette étude :

- ⇨ Période pré-incendie : les 32 premiers individus (semaine 1 à 32 inclues) du jeu de données
- ⇨ Période post-incendie : les 26 derniers individus (semaine 180 à 205 inclues) du jeu de données

La sélection et le filtrage des données correspondant à ces deux périodes sont effectués directement à partir du logiciel SAS Entreprise Guide.

Pour procéder à l'estimation des coefficients du modèle de régression simple, on utilise également le logiciel SAS Entreprise Guide par le biais de l'outil de régression linéaire avec une seule variable quantitative. On obtient alors les résultats suivants :

Période pré-incendie

Nb d'observations lues	32
Nb d'obs. utilisées	32

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	180066717	180066717	1424.09	<.0001
Erreur	30	3793289	126443		
Total sommes corrigées	31	183860006			

Root MSE	355.58820	R carré	0.9794
Moyenne dépendante	3636.56250	R car. ajust.	0.9787
Coeff Var	9.77814		

Résultats estimés des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	16.22868	114.69507	0.14	0.8884
A-Park	1	0.69349	0.01838	37.74	<.0001

Période post-incendie

Nb d'observations lues	26
Nb d'obs. utilisées	26

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	1	426295375	426295375	1299.04	<.0001
Erreur	24	7875875	328161		
Total sommes corrigées	25	434171250			

Root MSE	572.85380	R carré	0.9819
Moyenne dépendante	10308	R car. ajust.	0.9811
Coeff Var	5.55760		

Résultats estimés des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	459.45488	295.43338	1.56	0.1330
A-Park	1	0.97008	0.02692	36.04	<.0001

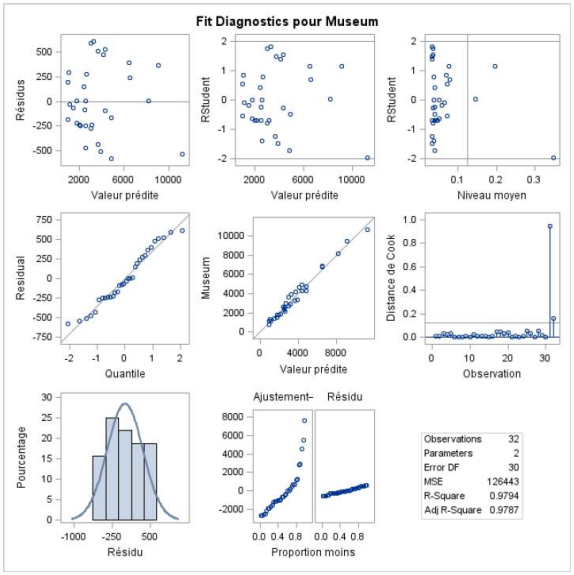
Les deux modèles utilisés confirment qu'il y a bien une liaison linéaire significative entre les fréquentations respectives du musée et du parc d'attraction pour chacun des deux échantillons observés. En effet, la valeur de la probabilité associée à un test de student pour chacun des deux modèles observés est inférieure à $\alpha = 5\%$, ce qui implique un rejet de l'hypothèse nulle de nullité du coefficient directeur de la droite de régression (slope) dans les deux cas.

Pour chacun des deux modèles, la qualité de la régression est globalement très bonne, bien que légèrement supérieure pour la période post-incendie ($R^2_{post} > R^2_{pre}$). En effet, 98,19% de la variance totale est expliquée par le modèle de régression sur cette période, contre 97,94% pour la période précédant l'incendie. Les données observées sur la période pré-incendie semblent toutefois indiquer une variance globale plus faible que sur la période post-incendie ($\hat{\sigma}_{pre} < \hat{\sigma}_{post}$). Si l'on considère les résultats estimés des paramètres, on observe que l'ordonnée à

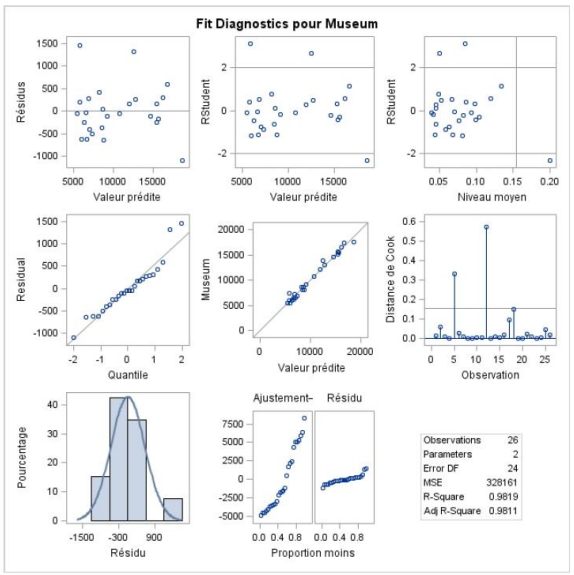
l'origine (l'intercept) n'est significative dans aucun des deux cas de figure car sa valeur de probabilité est respectivement supérieure à $\alpha = 5\%$.

On procède ensuite à une analyse des résidus studentisés qui viennent confirmer les hypothèses de normalité des erreurs, d'absence d'hétéroscédasticité et d'indépendance des termes d'erreur pour les deux modèles comparés. La droite de Henry et l'histogramme des effectifs peuvent également être des outils diagnostiques graphiques intéressants.

Période pré-incendie



Période post-incendie



Soient m et p les estimations de fréquentation observées respectives, précédant et succédant à l'incendie. On obtient alors les droites de régression suivantes :

- ⇨ Modèle pré-incendie : m = 0,69349 * A-Park
- ⇨ Modèle post-incendie : p = 0,97008 * A-Park

Le coefficient directeur de la droite de régression correspondant aux données pré-incendie est plus faible que celui correspondant aux données post-incendie. En d'autres termes, la fréquentation totale estimée par l'échantillon pré-incendie sera moins important que celle estimée par l'échantillon post-incendie. La compagnie d'assurance aura ainsi tout intérêt à sélectionner le premier modèle (période pré-incendie) au détriment du second (période post-incendie).

Une fois le modèle choisi, on peut alors estimer la fréquentation totale pour la période de fermeture du musée. Cette dernière s'élève à 783 177 personnes entre la semaine 33 et la semaine 179, ce qui représente une moyenne de 5 328 personnes par semaine pour les 147 semaines observées.

2. Régression Multiple

Une société souhaite faire l'acquisition d'un nouveau véhicule qui sera utilisé pour les déplacements de son équipe commerciale. Une grande partie des trajets étant effectuée sur l'autoroute, la société cherche donc à estimer la consommation d'essence sur l'autoroute d'un véhicule donné en fonction de ses caractéristiques techniques.

On cherche donc à identifier le meilleur modèle de régression multiple pour prédire la consommation d'essence sur l'autoroute (variable dépendante) à partir de trois caractéristiques sélectionnées en amont (variables quantitatives). L'objectif de l'étude est donc de procéder à l'estimation des coefficients du modèle de régression et déterminer la significativité des facteurs mesurés.

Le jeu de données choisi pour la société comprend initialement un grand nombre de variables quantitatives et qualitatives pour un échantillon de 93 individus, correspondant ici à 93 modèles de véhicules différents. Toutefois, seules quatre variables quantitatives sont retenues pour cette étude : la consommation d'essence sur l'autoroute (MPG Highway), l'empattement (Wheelbase), la puissance (Horspower) et le poids (Weight). La sélection et le filtrage des données observées sont effectués directement à partir du logiciel SAS Entreprise Guide.

Pour procéder à l'estimation des coefficients du modèle de régression multiple, on utilise également le logiciel SAS Entreprise Guide par le biais de l'outil de régression linéaire avec trois variables quantitatives. On obtient alors les résultats suivants :

Nb d'observations lues	93
Nb d'obs. utilisées	93

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	3	1821.68191	607.22730	68.10	<.0001
Erreur	89	793.62992	8.91719		
Total sommes corrigées	92	2615.31183			

Root MSE	2.98617	R carré	0.6965
Moyenne dépendante	29.08602	R car. ajust.	0.6863
Coeff Var	10.26667		

Résultats estimés des paramètres

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	26.26985	7.65972	3.43	0.0009
Horsepower	1	0.01155	0.01003	1.15	0.2526
Wheelbase	1	0.35628	0.10605	3.36	0.0012
Weight	1	-0.01168	0.00159	-7.35	<.0001

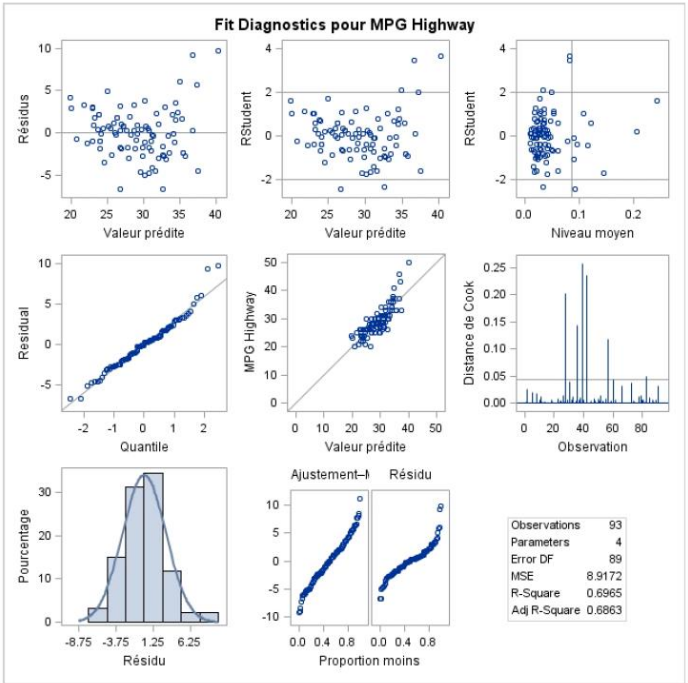
Au regard de ces premiers résultats, la qualité de la régression semble tout à fait acceptable puisque 68,63% de la variance totale est expliquée par le modèle de régression. L'observation du R^2 ajusté plutôt que du R^2 prévu est préférable ici car il tient compte du nombre de prédicteurs dans le modèle.

Le modèle utilisé confirme bien qu'il y a une liaison linéaire significative entre les trois variables quantitatives choisies et la variable dépendante. En effet, la valeur de la probabilité associée à un test de Fisher pour le modèle observé est inférieure à α = 5%, pour une valeur de la statistique F égale à 68,10.

Cela qui implique donc un rejet de l'hypothèse nulle de nullité du coefficient directeur de la droite de régression (slope).

De même, l'étude de la valeur de probabilité associée à un test de student pour chacune des variables retenues au sein du modèle permet de conclure sur leur significativité. En effet, le poids (Weight) et l'empattement (Wheelbase) ont des valeurs de probabilité inférieures à α = 5%, et doivent ainsi être conservés au sein du modèle. De la même manière, l'ordonnée à l'origine (l'intercept) est également retenue. En revanche, la puissance (Horsepower) est considérée comme non-significative, l'hypothèse nulle de nullité du paramètre n'étant pas rejetée.

On procède ensuite à l'étude des résidus permettant de vérifier empiriquement les hypothèses initiales nécessaires au bien-fondé du modèle.



L'analyse des résidus studentisés permet de conclure quant à la normalité des erreurs. L'étude graphique de la droite de Henry et de l'histogramme des effectifs peuvent également être des outils diagnostiques utiles.

On procède ensuite à un test D de Durbin-Watson afin de vérifier l'indépendance des termes d'erreur. Etant proche de 2, la statistique du test obtenue (1,607) permet de conclure à l'absence d'autocorrélation.

Corrélation des valeurs estimées				
Variable	**Intercept**	**Horsepower**	**Wheelbase**	**Weight**
Intercept	1.0000	-0.4059	-0.9742	0.7532
Horsepower	-0.4059	1.0000	0.4766	-0.7348
Wheelbase	-0.9742	0.4766	1.0000	-0.8701
Weight	0.7532	-0.7348	-0.8701	1.0000

L'étude de la matrice des corrélations entre les variables observées permet de conclure à la non-multicolinéarité.

Les hypothèses du modèle étant maintenant vérifiées, on peut alors en déduire le modèle de régression Y à partir duquel la société pourra estimer sa consommation d'essence sur l'autoroute pour n'importe quel modèle de véhicule : MPG Highway = 26,26985 + (0,35628 * Wheelbase) – (0,01168 * Weight)

La société décide d'opter pour le modèle de véhicule ATS-110 2019 avec les caractéristiques suivantes :

- Weight = 3 800
- Horsepower = 300
- Wheelbase = 100

Grâce au modèle précédemment proposé, la société peut alors estimer qu'elle devra prévoir la consommation d'essence suivante : 26,26985 + (0,35628 * 100) – (0,01168 * 3 800) = 106.28185 MPG

3. Analyse de la variance à plusieurs facteurs

Un hôpital souhaite effectuer une étude concernant un test de tension nerveuse sur ses patients afin d'estimer les caractéristiques des personnes susceptibles d'atteindre plus rapidement un niveau prédéfini de tension nerveuse pouvant déclencher des risques pathologiques importants.

L'objectif est d'identifier les niveaux optimaux de ces caractéristiques qui permettent d'atteindre le niveau de réponse souhaité en moins de temps possible. Pour ce faire, on procède à l'analyse des résultats de la variance afin de vérifier le caractère significatif des facteurs sélectionnés sur le niveau de tension nerveuse.

L'échantillon comprend n_{total} = 36 individus pour quatre variables mesurées. Ces variables sont composées de trois facteurs identifiés en amont par le corps médical (variables de classification), et de la réponse correspondant au temps en minutes avant que le sujet observé n'atteigne le niveau de tension nerveuse prédéfini (variable dépendante quantitative).

	Facteurs	Niveaux des facteurs	Taille de l'échantillon
A	Corpulence (body fat)	Petite (low)	$n_A = 2$
		Importante (high)	
B	Sexe (gender)	Masculin (male)	$n_B = 2$
		Féminin (female)	
C	Fumeur (smoking)	Jamais (none)	$n_C = 3$
		Un peu (light)	
		Beaucoup (heavy)	

Pour procéder à l'estimation des paramètres du modèle, on utilise le logiciel SAS Entreprise Guide par le biais de l'outil d'ANOVA à plusieurs facteurs. Les interactions d'ordre trois ou supérieure à trois sont exclues de l'analyse. On obtient alors les résultats suivants :

Le modèle utilisé confirme l'existence d'une liaison linéaire significative entre les trois variables de classification choisies et la variable dépendante. En effet, la valeur de probabilité associée au test de Fisher pour le modèle observé est inférieure à α = 5%, pour une valeur de la statistique F égale à 165,24. Cela implique donc un rejet de l'hypothèse nulle de nullité simultanée des coefficients du modèle.

De la même façon, il existe une liaison linéaire significative entre chacun des facteurs et la réponse observée. En effet, on constate que les valeurs de probabilité associées au test de Fisher sont toutes inférieures à α = 5%.

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	10	14670.16667	1467.01667	165.24	<.0001
Erreur	26	230.83333	8.87821		
Total sommes non cor	36	14901.00000			

R-carré	Coef de var	Racine MSE	minutes Moyenne
0.865349	15.56847	2.979632	19.13889

Source	DDL	Type I SS	Carré moyen	Valeur F	Pr > F
Constante	1	13186.69444	13186.69444	1485.29	<.0001
body fat	1	702.25000	702.25000	79.10	<.0001
gender	1	210.25000	210.25000	23.68	<.0001
smoking	2	343.05556	171.52778	19.32	<.0001
gender*smoking	2	21.50000	10.75000	1.21	0.3142
body fat*smoking	2	204.16667	102.08333	11.50	0.0003
body fat*gender	1	2.25000	2.25000	0.25	0.6189

Source	DDL	Type III SS	Carré moyen	Valeur F	Pr > F
Constante	1	13186.69444	13186.69444	1485.29	<.0001
body fat	1	702.25000	702.25000	79.10	<.0001
gender	1	210.25000	210.25000	23.68	<.0001
smoking	2	343.05556	171.52778	19.32	<.0001
gender*smoking	2	21.50000	10.75000	1.21	0.3142
body fat*smoking	2	204.16667	102.08333	11.50	0.0003
body fat*gender	1	2.25000	2.25000	0.25	0.6189

En d'autres termes, les trois facteurs A, B et C influent de manière significative sur la réponse Y.

Il semble que seule l'interaction AC soit également significative. Pour améliorer la précision, il faut donc procéder de nouveau à l'analyse de la variance en excluant cette fois-ci les interactions non-significatives AB et BC.

Une fois la seconde analyse effectuée, la somme des carrés obtenue est alors inférieure à celle précédemment obtenue pour une valeur de la statistique F égale à 27,71.

Source	DDL	Somme des carrés	Carré moyen	Valeur F	Pr > F
Modèle	6	1459.722222	243.287037	27.71	<.0001
Erreur	29	254.583333	8.778736		
Total sommes corrigé	35	1714.305556			

R-carré	Coef de var	Racine MSE	minutes Moyenne
0.851495	15.48101	2.962893	19.13889

Source	DDL	Type I SS	Carré moyen	Valeur F	Pr > F
body fat	1	702.2500000	702.2500000	79.99	<.0001
gender	1	210.2500000	210.2500000	23.95	<.0001
smoking	2	343.0555556	171.5277778	19.54	<.0001
body fat*smoking	2	204.1666667	102.0833333	11.63	0.0002

Source	DDL	Type III SS	Carré moyen	Valeur F	Pr > F
body fat	1	702.2500000	702.2500000	79.99	<.0001
gender	1	210.2500000	210.2500000	23.95	<.0001
smoking	2	343.0555556	171.5277778	19.54	<.0001
body fat*smoking	2	204.1666667	102.0833333	11.63	0.0002

La valeur de probabilité associée au test de Fisher pour le modèle observé est quant à elle toujours inférieure à $\alpha = 5\%$. On obtient alors le modèle suivant :

Minutes = 32,92 + [-14,67 0,00]*Corpulence + [-4,83 0,00]*Sexe + [-13,33 -7,50 0,00]*Fumeur

$$+ \begin{bmatrix} 11{,}67 & 5{,}83 & 0{,}00 \\ 0{,}00 & 0{,}00 & 0{,}00 \end{bmatrix} \text{*Corpulence*Fumeur}$$

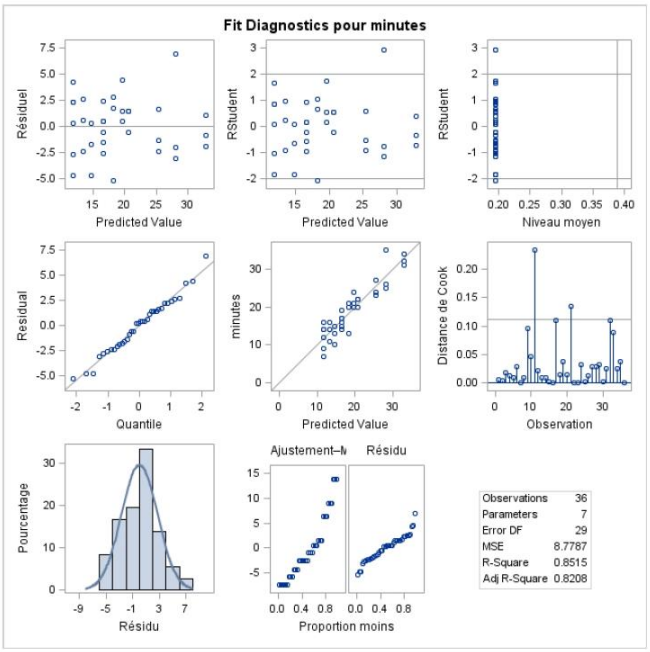
On procède alors à l'étude des résidus permettant de vérifier empiriquement les hypothèses initiales nécessaires au bien-fondé du modèle.

L'analyse des résidus studentisés permet de conclure à :

- ⇨ La normalité des erreurs
- ⇨ L'absence d'hétéroscédasticité
- ⇨ L'indépendance des termes d'erreur

L'étude graphique de la droite de Henry et de l'histogramme des effectifs peuvent également être des outils diagnostiques utiles.

On s'intéresse ensuite à l'estimation des paramètres des niveaux des facteurs afin de déterminer la configuration optimale



.

Les caractéristiques des personnes susceptibles d'atteindre plus rapidement un niveau prédéfini de tension nerveuse sont donc les suivantes :

- ⇨ Corpulence (body fat) : importante (high)
- ⇨ Sexe (gender) : féminin (female)
- ⇨ Fumeur (smoking) : beaucoup (heavy)

Paramètre	Estimation		Erreur type	Valeur du test t	Pr > |t|
Constante	32.91666667	B	1.30651306	25.19	<.0001
body fat high	-14.66666667	B	1.71062714	-8.57	<.0001
body fat low	0.00000000	B	.	.	.
gender female	-4.83333333	B	0.98763104	-4.89	<.0001
gender male	0.00000000	B	.	.	.
smoking heavy	-13.33333333	B	1.71062714	-7.79	<.0001
smoking light	-7.50000000	B	1.71062714	-4.38	0.0001
smoking none	0.00000000	B	.	.	.
body fat*smoking high heavy	11.66666667	B	2.41919210	4.82	<.0001
body fat*smoking high light	5.83333333	B	2.41919210	2.41	0.0225
body fat*smoking high none	0.00000000	B	.	.	.
body fat*smoking low heavy	0.00000000	B	.	.	.
body fat*smoking low light	0.00000000	B	.	.	.
body fat*smoking low none	0.00000000	B	.	.	.

4. Plans hiérarchiques & Modèles à effets aléatoires

La société Paritexo vient de remporter un appel d'offre décisif concernant la commercialisation d'un nouveau type de radar plus performant, dont l'objectif est de pouvoir contrôler la vitesse des véhicules motorisés sur les voies de circulation rapide et sur les autoroutes. La nouveauté de ce radar réside dans sa capacité à mesurer les excès de vitesse de moins de 20km/h.

Afin de vérifier le niveau de performance de l'outil, la société Paritexo souhaite donc effectuer une étude de la fidélité avec pour objectif une incertitude de moins de 5% quant à la valeur mesurée sur les autoroutes.

La construction du schéma expérimental et l'enregistrement des mesures obtenues sont effectués par une équipe dédiée au sein de la société Paritexo. Cette dernière fait le choix de procéder à cinq séances (k = 5) composées pour chacune de quatre essais (n = 4).

Essais	Séances				
	1	2	3	4	5
1	131	128	139	132	131
2	140	135	132	141	135
3	124	129	126	128	129
4	135	137	131	127	136

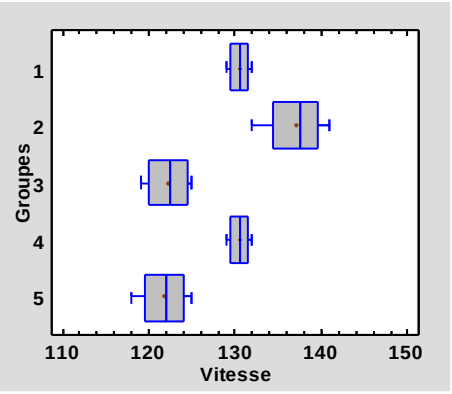
Les mesures ont été effectuées avec brio, et n'ont pas données lieu à de quelconques données manquantes. Ce schéma expérimental peut être également transposé sous la forme du plan d'expérience suivant :

Groupes	Répétabilité	Vitesse observée
1	1	131
1	2	129
1	3	132
1	4	130
2	1	137
2	2	138
2	3	141
2	4	132
3	1	121
3	2	125
3	3	124
3	4	119
4	1	130
4	2	131
4	3	129
4	4	132
5	1	121
5	2	123
5	3	118
5	4	125

La variable « Groupes » correspond aux différentes séances de mesure effectuées, tandis que la variable « Répétabilité » correspond au nombre d'essais réalisés lors de chacune des séances.

La première étape de l'étude consiste à déterminer la composante de répétabilité S_r^2 pour la vitesse observée. Les écarts-types sont donc calculés pour chacun des groupes.

Groupes	Comptage	Moyenne	Ecart-type
1	4	130,5	1,29099
2	4	137,0	3,74166
3	4	122,25	2,75379
4	4	130,5	1,29099
5	4	121,75	2,98608
Total	20	128,4	6,32788



Afin de vérifier s'il y a une différence statistiquement significative entre les écarts-types de chacun des groupes au niveau de confiance 95,0%, on procède au test de Levene.

Tests des variances	
Test de Levene	1,16518
Probabilité	0,3651

La valeur de probabilité obtenue est supérieure à 0,05 ce qui signifie que l'on ne rejette pas l'hypothèse nulle correspondant à une égalité statistiquement significative entre les écarts-types avec un niveau de confiance de 95,0%.

Etant donnée que la variance intergroupe est significative, la variance de fidélité intermédiaire peut alors être estimée à partir de ses composants pour le modèle à effets aléatoires :

Source	Somme des carrés	Degrés de liberté	Carré moyen	Composant de la variance	Contribution
Groupes	659,3	4	164,825	39,5146	85,38%
Répétabilité	101,5	15	6,76667	6,76667	14,62%
Total	760,8	19			100,00%

Le tableau de l'analyse de la variance ci-dessus décompose la variance de la vitesse observée en deux composants, un pour chacun des facteurs. A l'exception du premier, chaque facteur est imbriqué dans le facteur qui précède. Le but d'une telle analyse est d'estimer le montant de la variabilité expliquée par chaque facteur, appelé composant de la variance.

Dans ce cas, le facteur contribuant le plus à la variance est le facteur « Groupes » correspondant aux différentes séances. Sa contribution représente 85,3792% du total de la variance de la vitesse observée.

L'essentielle de la variabilité (85,3792% du total de la variance de la vitesse observée) est issue de la variabilité intergroupe, ce qui signifie que les répétitions inter-séances (k) apportent davantage de précision que les répétitions intra-séances (n). Aussi, la société Paritexo peut procéder à davantage de séances (k' > k) afin d'améliorer la précision.

La seconde étape de l'étude consiste à calculer la variance de fidélité intermédiaire S_R^2 pour la vitesse observée à partir de l'égalité suivante :

$$\begin{cases} S_R^2 = S_r^2 + S_g^2 = 6{,}76667 + 39{,}5146 = 46{,}28127 \\ S_R = \sqrt{{S_R}^2} = \sqrt{46{,}28127} \approx 6{,}8030 \end{cases}$$

L'écart-type de fidélité intermédiaire est donc : $S_R \approx 6{,}8030$

Il est alors possible de calculer l'intervalle de confiance des résultats :

$$IC_R = \text{vitesse limite} \pm t_{\alpha,\upsilon} \sqrt{\frac{{S_g}^2}{k} + \frac{{S_r}^2}{kn}} = 130 \pm 2{,}1 \sqrt{\frac{39{,}5146}{k} + \frac{6{,}76667}{kn}}$$

Pour k = 1 et n = 1, on obtient l'intervalle de confiance suivant : $IC_R = 130 \pm 14{,}286 = [115{,}714 ; 144{,}286]$

En conclusion, pour une vitesse limite de 130 km/h, les résultats de l'évaluation de la fidélité sont les suivants :

- Ecart-type de répétabilité : $S_r \approx 2{,}601$ avec 15 degrés de liberté
- Ecart-type de fidélité intermédiaire : $S_R \approx 6{,}803$ avec 19 degrés de liberté

5. 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 : p-value = 0,0078 < 0,01
- Le temps a un effet hautement significatif : p-value < 0,001
- L'interaction entre le produit et le temps a un effet hautement significatif : 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 + (Produit A)*(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.

6. 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] :

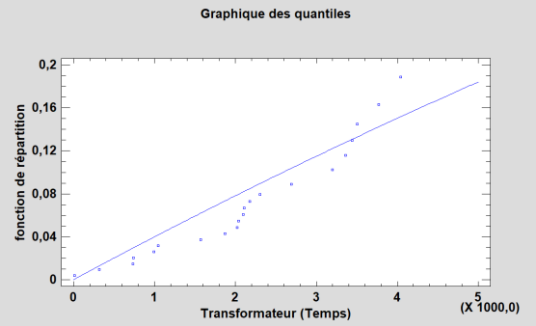
Observation	Temps	Statut	Nombre de présents	Survie cumulée	Erreur-type	Risque cumulé
1	10	DEFAILLANCE	179	0,9944	0,0055	0,0056
2	314	DEFAILLANCE	178	0,9889	0,0078	0,0112
3	730	DEFAILLANCE	177	0,9833	0,0095	0,0168
4	740	DEFAILLANCE	176	0,9778	0,011	0,0225
5	990	DEFAILLANCE	175	0,9722	0,0122	0,0282
6	1046	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.

7. 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