Projet de Statistique for Data Science

AUTRICE

Marie-Ange DIENG

2025-02-04

Table des matières

1 Introduction

1.1 Contexte et Objectif

L'analyse de survie est une technique statistique utilisée pour analyser et modéliser le temps jusqu'à la survenue d'un événement d'intérêt, comme un décès, une rechute de maladie, ou l'arrêt d'un processus. Elle est largement utilisée dans des domaines comme la médecine, l'épidémiologie, l'ingénierie et la recherche sociale. Contrairement aux analyses de données classiques, l'analyse de survie prend en compte le temps jusqu'à l'événement, ainsi que les individus censurés, c'est-à-dire ceux dont l'événement d'intérêt n'est pas observé pendant la période d'étude.

Dans ce projet, l'objectif principal est d'appliquer un modèle de régression de Cox[1] pour analyser la survie d'individus en fonction de plusieurs facteurs explicatifs. Plus précisément, nous cherchons à déterminer quels facteurs, parmi des caractéristiques telles que l'âge, la consommation d'alcool et l'usage de drogues, influencent la durée de survie d'un individu et leur probabilité d'expérimenter un événement (par exemple, un décès). Le modèle de Cox est un modèle semi-paramétrique qui permet d'étudier la relation entre plusieurs covariables et le temps de survie, tout en n'exigeant pas de spécification explicite de la forme de la fonction de risque de base. Il permet ainsi une analyse flexible et puissante, même en l'absence d'une distribution paramétrique spécifique pour les données.

Le modèle de Cox est une méthode particulièrement populaire dans l'étude de la survie dans des contextes médicaux et cliniques, où les chercheurs souhaitent étudier l'impact de divers facteurs sur le temps avant qu'un événement ne se produise, sans faire d'hypothèses strictes sur la forme de la distribution des données de survie.

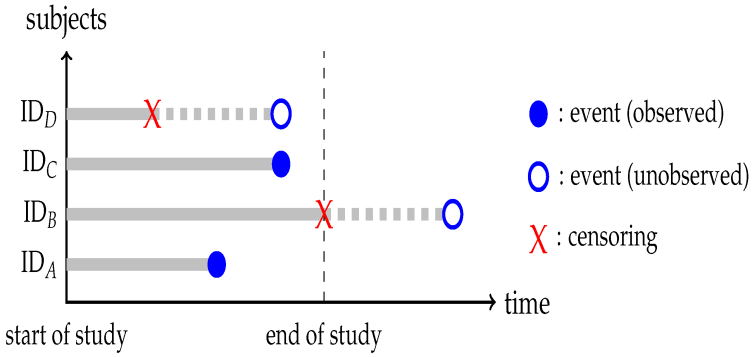

FIGURE 1 – Représentation d'un exemple de l'analyse de survie

1.2 Description du jeu de données

Pour mener cette analyse, nous avons utilisé un dataset[2] contenant des informations sur un ensemble d'individus. Ce dataset, d'une taille de 1 658 observations, inclut plusieurs variables liées à

la durée de suivi des individus et à leur état de santé. Voici une description détaillée des principales variables de ce dataset :

- **death :** Variable binaire indiquant si l'individu est décédé pendant la période d'observation (1 = Décédé, 0 = Vivant).
- **agein :** L'âge de l'individu au début de la période d'observation.
- **alcohol :** Niveau de consommation d'alcool de l'individu, où 1 indique une faible consommation d'alcool et 2 une forte consommation.
- **drug.use :** Indicateur binaire de l'utilisation de drogues illicites (0 = Non, 1 = Oui).
- **timein :** La date de début de l'observation pour chaque individu.
- **timeout :** La date de fin de l'observation ou la date de l'événement (décès ou dernière observation).
- **timebth :** La date de naissance de l'individu, utilisée ici pour calculer l'âge.
- **duree :** Durée de suivi de l'individu en jours, *calculée* à partir des dates timein et timeout. Dans le cadre de ce projet, la variable death a été utilisée comme variable cible pour modéliser la durée de survie, et les variables explicatives incluent agein, alcohol et drug.use. Le temps jusqu'à l'événement, mesuré en jours, constitue la variable de temps d'intérêt, duree, qui sera utilisée pour analyser les facteurs de risque influençant la survie des individus.

1.3 Méthode de l'analyse de survie avec le modèle de Cox

Le modèle de Cox, ou modèle de risques proportionnels de Cox, est un modèle statistique qui permet de déterminer l'effet de plusieurs variables explicatives sur le risque de survenue d'un événement en fonction du temps. Contrairement aux modèles paramétriques (comme le modèle de Weibull), le modèle de Cox ne fait aucune hypothèse sur la forme de la fonction de risque de base, ce qui en fait un outil flexible et robuste pour analyser des données de survie. La fonction de hazard de Cox à pour formule : La fonction de hazard de Cox à pour formule :

$$h(t|X) = h_0(t) \cdot \exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n\right)$$

avec :
$h(t|X)$: La fonction de hazard à un moment t pour les covariables $X = (X_1, X_2, \ldots, X_n)$
$h_0(t)$: La fonction de risque de base, qui représente le risque que l'événement se produise quand toutes les covariables sont nulles.
$\beta_1, \beta_2, \ldots, \beta_n$: Les coefficients correspondants à chaque covariable. Ce sont les rapports de cotes ou hazard ratios.

Dans ce modèle, le risque relatif d'un événement pour un individu est estimé en fonction de ses caractéristiques (covariables). L'hypothèse de proportionalité des risques stipule que le rapport des risques entre deux individus reste constant au cours du temps, bien que leurs caractéristiques varient. Cette hypothèse est essentielle pour l'interprétation des résultats du modèle de Cox, notamment les

rapports de cotes ou hazard ratios (HR), qui indiquent dans quelle mesure un changement dans une variable explicative (par exemple, une augmentation de l'âge) affecte le risque d'un événement.

L'un des atouts majeurs du modèle de Cox est qu'il ne nécessite pas que la forme de la fonction de risque de base soit spécifiée, ce qui permet de se concentrer sur les relations entre les covariables et les risques relatifs. Cependant, le modèle repose sur certaines hypothèses importantes, notamment la proportionalité des risques, qui sera vérifiée dans le cadre de l'analyse.

1.4 Objectifs spécifiques du projet

L'objectif principal de ce projet est d'appliquer le modèle de Cox pour identifier les facteurs associés à la survie dans un échantillon d'individus, en particulier l'âge, la consommation d'alcool et l'usage de drogues. Plus précisément, ce projet vise à :

1. Estimer l'impact des covariables (âge, alcool, usage de drogues) sur le temps de survie des individus.
2. Examiner l'effet de ces variables sur la probabilité de survenue de l'événement (par exemple, un décès), tout en contrôlant les autres facteurs.
3. Estimer la fonction de risque cumulée de base et analyser comment elle varie en fonction des covariables.
4. Discuter des résultats et de leur signification pratique, en particulier en ce qui concerne les politiques de santé publique liées à la consommation d'alcool et de drogues.

2 Méthodologie

2.1 Préparation et nettoyage des données

Avant de procéder à l'analyse, il est nécessaire de préparer et de nettoyer les données. Cela inclut plusieurs étapes importantes pour garantir que les données sont au bon format et prêtes pour l'analyse.

2.1.1 Chargement des données

Les données sont chargées à partir d'un fichier CSV à l'aide de la fonction read.table(). Voici le code utilisé pour cela :

```
1 data <- read.table(file.choose(), header=TRUE, dec="." , sep=',', na.
    strings="999")
```

2.1.2 Conversion des colonnes de dates

Les colonnes timein, timeout, et timebth contiennent des informations de dates sous forme de chaîne de caractères. Ces dates doivent être converties en format Date dans R afin de pouvoir être manipulées correctement dans les analyses suivantes.

```
timein <- as.Date(timein, format = "%d/%m/%Y")
timeout <- as.Date(timeout, format = "%d/%m/%Y")
timebth <- as.Date(timebth, format = "%d/%m/%Y")
```

2.1.3 Calcul de la durée de suivi en jours

La durée de suivi représente le temps écoulé entre timein et timeout. Cette variable est cruciale pour l'analyse de survie, car elle représente le temps jusqu'à l'événement d'intérêt (ici, le décès). Elle est stockée en 'jours'.

```
data$duree <- as.numeric(difftime(data$timeout, data$timein, units = "
    days"))
```

2.2 Création de l'objet de survie

Une fois les données préparées, il est nécessaire de créer un objet de survie. Dans le modèle de Cox, cet objet permet de spécifier les informations sur le temps et l'événement à étudier. Le modèle de survie est créé avec la fonction Surv() de la bibliothèque survival, qui prend en entrée deux arguments : le temps (ici, la variable duree) et l'événement (ici, la variable death).

```
install.packages("survival")
library(survival)
surv_obj <- Surv(time = data$duree, event = data$death)
```

— time = data$duree : La variable duree représente le temps de suivi, c'est-à-dire le nombre de jours jusqu'à l'événement ou la censure.

— event = data$death : La variable death représente l'événement d'intérêt (1 si l'événement s'est produit, 0 s'il n'a pas eu lieu).

2.3 Ajustement du modèle de Cox

Une fois que l'objet de survie est créé, on peut ajuster le modèle de Cox. Le modèle de Cox est utilisé pour analyser l'influence de plusieurs variables indépendantes (covariables) sur le temps de survie.

```
cox_model <- coxph(surv_obj ~ agein + alcohol + drug.use, data = data)
```

— coxph() : Cette fonction ajuste le modèle de Cox aux données. Le modèle estime l'effet des variables explicatives sur la probabilité de l'événement (décès) en fonction du temps.

— La formule survobj = agein + alcohol + drug.use spécifie que nous souhaitons examiner l'effet des variables agein, alcohol et drug.use sur la durée de survie, en contrôlant pour d'autres facteurs.

2.4 Estimation de la fonction de risque cumulée de base

La fonction de risque cumulée de base est une estimation de la fonction de risque pour un individu dont toutes les covariables sont égales à zéro. Cela permet de mieux comprendre la structure temporelle du risque.

```
base_hazard <- basehaz(cox_model, centered = FALSE)
plot(base_hazard$time, base_hazard$hazard, type = "l", main = "Baseline
    Hazard Function")
```

3 Résultats

3.1 Résumé des résultats du modèle de Cox

La sortie du modèle nous fournit les estimations des coefficients (β), les rapports de risques ($\exp(\beta)$), les erreurs standards des coefficients ($se(\beta)$), les valeurs de z-statistique et les p-values correspondantes pour chaque variable explicative. Voici un résumé des résultats du modèle :

Variable	**Coefficient** (β)	**Hazard Ratio** ($\exp(\beta)$)	**Erreur Standard**	z**-statistique**	p**-value**
`agein`	0.085992	1.089798	0.008341	10.309	$< 2 \times 10^{-16}$
`alcohol`	0.249900	1.283896	0.108566	2.302	0.0213
`drug.use`	0.088089	1.092086	0.102303	0.861	0.3892

TABLE 1 – Résumé des résultats du modèle de Cox.

3.2 Interprétation des coefficients et des rapports de risques

1. Variable *agein* (âge au début de l'étude) : Le coefficient estimé pour agein est de 0.085992, ce qui signifie que, toutes choses égales par ailleurs, l'augmentation de l'âge de 1 an est associée à une augmentation du risque de décès d'environ $8,6\%$. Le rapport de risque ($exp(\beta)$) est égal à 1.089798 (>1), donc celà comfirme que l'age à un impact sur le risque de décès. De plus le p-value est $< 2 \times 10^{-16} (< 0,05)$, ce qui confirme que l'association age-risque de décès est statistiquement très significative.

2. Variable *alcohol* (consommation d'alcool) : Le coefficient estimé pour alcohol est de 0.249900. Cela signifie qu'une augmentation du niveau de consommation d'alcool (par exemple, d'une faible consommation à une consommation élevée) est associée à une augmentation du risque de décès d'environ 28.4%. Le rapport de risque ($exp(\beta)$) est égal à 1.283896. (>1), donc celà comfirme que le niveau de consommation d'alcool à un impact sur le risque de décès. De plus le p-value est 0.0213 (<0,05), ce qui confirme que l'association age-risque de décès est statistiquement significative.
3. Variable *drug.use* (usage de drogues) : Le coefficient estimé pour drug.use est de 0.088089.Cela signifie qu'une personne qui utilise des drogues illicites a un risque de décès supérieur de 9.2% par rapport à une personne ne consommant pas de drogues.Le rapport de risque ($exp(\beta)$) est égal à 1.092086. (>1), donc celà comfirme que la consommation de drogue à un impact sur le risque de décès. Cependant, le p-value de 0.3892 indique que cette relation n'est pas statistiquement significative, ce qui suggère qu'aucune preuve claire n'existe pour l'association entre l'usage de drogues et le risque de décès dans cette analyse. Il se pourrait que d'autres facteurs non inclus dans le modèle aient un effet plus fort.

3.3 Estimation de la fonction de risque cumulée de base

Le modèle de Cox ne fournit pas directement la fonction de risque de base $h_0(t)$ mais nous pouvons l'estimer à l'aide de la fonction $basehaz()$ de R. Cette fonction donne la valeur de la fonction de risque cumulée de base dans le temps, qui est indépendante des effets des covariables. Nous avons obtenu les valeurs suivantes : Ces données nous on permis de générer la courbe suivante, qui montre

Hazard	Time
0.000000e+00	-10531
0.000000e+00	-10500
0.000000e+00	-10258
0.000000e+00	-10197
0.000000e+00	-10196
...	...
9.886637e-04	-549
9.886637e-04	-519
...	...

TABLE 2 – Données de hazard et de temps.

l'évolution de la fonction de risque cumulée de base avec le temps :

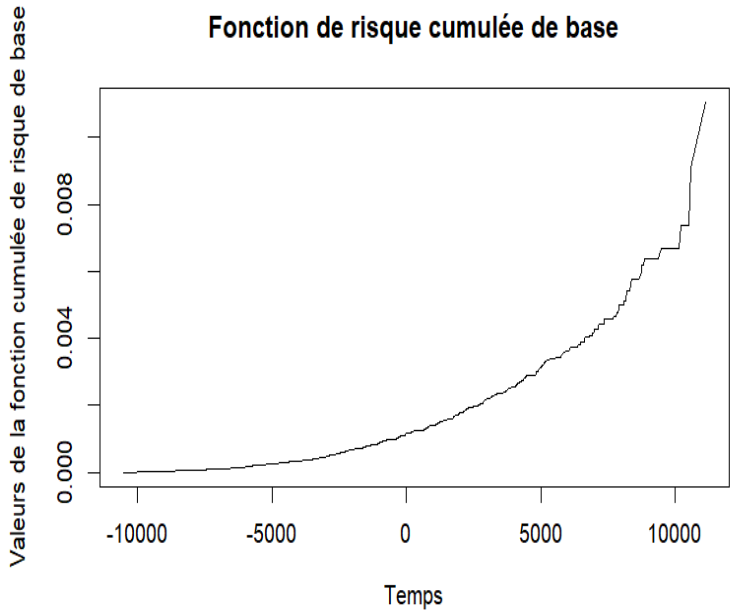

FIGURE 2 – Fonction de risque cumulée de base en fonction du temps

3.4 Test de l'hypothèse de proportionnalité des risques

Un des tests importants dans le modèle de Cox est l'hypothèse de proportionnalité des risques, selon laquelle les rapports de risques (hazard ratios) sont constants au fil du temps. Cette hypothèse peut être testée à l'aide de tests de Schoenfeld (ou tests de la proportionnalité des risques).

```
cox.zph(cox_model)
```

Les résultats de ce test indiquent si l'hypothèse de proportionnalité des risques est respectée. Tous les

Variable	Chisq	df	p-value
agein	1.756	1	0.19
alcohol	0.962	1	0.33
drug.use	1.157	1	0.28
GLOBAL	3.081	3	0.38

TABLE 3 – Résultats du test chi-carré pour les variables explicatives.

p-values sont supérieurs à 0.05. Le test est donc non significatif ce qui veut dire que l'hypothèse de proportionnalité des risques est valide.

4 Discussion

4.1 Rappel des résultats

Les résultats du modèle de Cox montrent que l'âge au début de l'étude et la consommation d'alcool sont des facteurs significativement associés à un risque accru de décès. En revanche, l'usage de drogues n'a pas montré d'association statistiquement significative avec le risque de décès dans cette population.

4.2 Comparaison avec la littérature existante

Les résultats de notre analyse sont cohérents avec plusieurs études précédentes qui ont examiné les facteurs de risque de décès en fonction de l'âge et de la consommation d'alcool. Par exemple, une étude de Jürgen Rehm [3] (2010) a montré que la consommation excessive d'alcool est un facteur de risque bien établi pour des événements de mortalité prématurée, y compris les maladies cardiovasculaires et les accidents.

En revanche, l'absence d'association significative entre l'usage de drogues et le risque de décès est moins fréquente dans la littérature, où l'on observe généralement une augmentation du risque, en particulier chez les jeunes adultes et les personnes ayant des comportements à risque. Il est donc possible que dans notre étude, la population soit plus homogène ou que d'autres facteurs, comme les comportements sociaux ou les comorbidités, n'aient pas été pris en compte.

4.3 Implications des résultats

Les résultats de cette analyse ont plusieurs implications, notamment pour la santé publique et les politiques de prévention. Le fait que l'âge et la consommation d'alcool soient des facteurs de risque significatifs suggère qu'il pourrait être utile de cibler ces facteurs dans les stratégies de prévention des décès prématurés. Par exemple :

— **Interventions ciblées pour les personnes âgées :** Des stratégies de dépistage et de prévention adaptées aux personnes âgées, qui pourraient inclure des examens de santé réguliers et des conseils sur le mode de vie, pourraient réduire les risques de décès associés à des conditions médicales liées à l'âge.

— **Réduction de la consommation d'alcool :** Les politiques de santé publique visant à réduire la consommation d'alcool pourraient avoir un impact significatif sur la réduction du risque de décès. Cela pourrait inclure des campagnes de sensibilisation, des restrictions sur la vente d'alcool, et des interventions pour traiter les dépendances à l'alcool.

4.4 Limitation de l'étude

Comme toute étude, cette analyse présente plusieurs limites qui doivent être prises en compte lors de l'interprétation des résultats :

1. **Manque de variables de confondance :** Bien que nous ayons inclus plusieurs variables dans le modèle, il est possible que des facteurs importants, tels que la condition physique, les antécédents médicaux, ou d'autres comportements de santé, n'aient pas été pris en compte. Cela peut entraîner un biais de confusion et limiter la capacité à tirer des conclusions causales nettes.
2. **Qualité des données** Les données utilisées dans cette analyse proviennent d'un jeu de données simulé, dérivé d'un ensemble de données anonymisé issu d'une étude de la London School of Hygiene and Tropical Medicine (LSHTM) portant sur l'association entre l'âge, la consommation d'alcool et le risque de décès. Bien que ces données soient anonymisées et issues d'une étude réelle, il existe toujours un risque d'erreurs liées à la collecte et à la manipulation des informations. Par exemple, bien que les données sur la consommation d'alcool et d'autres comportements de santé aient été collectées de manière rigoureuse, des biais peuvent encore exister, notamment en raison de la nature auto-déclarée de certaines variables, telles que la consommation d'alcool. Ces biais peuvent influencer la précision des estimations et limiter la capacité à tirer des conclusions définitives. Cependant, étant donné la nature anonymisée et les précautions prises dans la collecte des données, ces erreurs de mesure sont considérées comme relativement faibles.

4.5 Recommandations pour des études futures

Pour améliorer la compréhension des facteurs associés au risque de décès, plusieurs pistes peuvent être envisagées dans les futures études :

1. **Inclure d'autres variables :** Pour obtenir une image plus complète des facteurs influençant la survie, il serait utile d'inclure des variables supplémentaires telles que les antécédents médicaux, le statut socio-économique, l'activité physique, et des informations sur les comportements alimentaires.
2. **Utiliser des modèles plus sophistiqués :** Si l'hypothèse de proportionnalité des risques est violée, il pourrait être intéressant d'utiliser des modèles alternatifs, comme le modèle de Cox à effets non proportionnels ou même des approches plus modernes comme les modèles de survie à gradient boosting.
3. **Étudier des interactions entre variables :** L'examen des interactions entre les variables explicatives (par exemple, l'interaction entre l'âge et l'alcool) pourrait fournir des insights supplémentaires sur les effets combinés de ces facteurs sur le risque de décès.
4. **Longitudinalité des données :** De futures études pourraient également bénéficier de données longitudinales, où des mesures répétées sur les mêmes individus au fil du temps permettraient d'examiner l'évolution du risque et l'impact des changements dans les comportements (par exemple, une réduction de la consommation d'alcool).

5 Conclusion

L'objectif principal de cette étude était d'examiner les facteurs influençant le risque de décès en utilisant un modèle de Cox de risques proportionnels, basé sur un jeu de données comprenant des informations sur l'âge, la consommation d'alcool et l'usage de drogues. À travers cette analyse, plusieurs éléments clés ont émergé, contribuant à notre compréhension du risque de décès dans cette population. En conclusion, cette étude a fourni une analyse utile des facteurs de risque de décès à partir des variables disponibles, identifiant l'âge et la consommation d'alcool comme des déterminants clés du risque de décès prématuré. Bien que l'usage de drogues n'ait pas montré un effet statistiquement significatif dans ce contexte, d'autres recherches pourraient explorer plus en profondeur cette relation. Les résultats mettent en lumière l'importance de stratégies de prévention ciblées, notamment pour les personnes âgées et celles ayant une consommation élevée d'alcool, et soulignent la nécessité d'une compréhension plus nuancée des facteurs affectant la survie à long terme. Des recherches futures devront continuer à affiner ces modèles pour inclure davantage de facteurs de risque et mieux comprendre leurs interactions complexes.

Table des figures

Liste des tableaux

6 Annexe

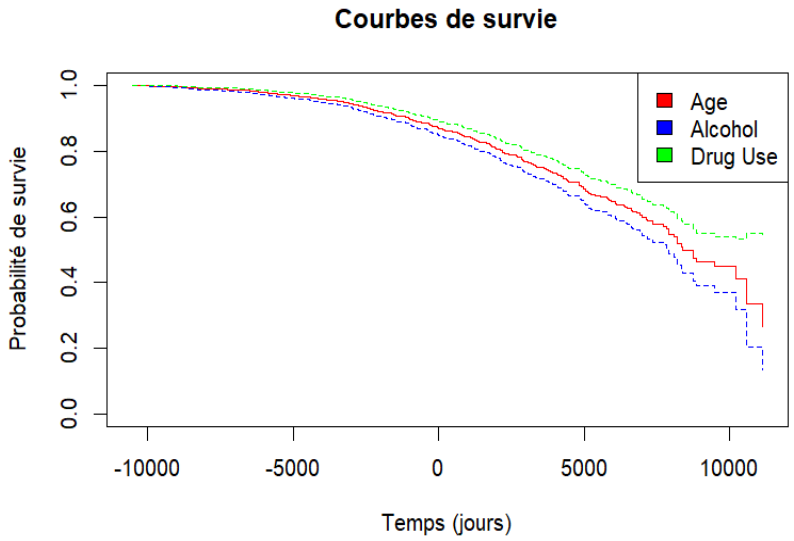

FIGURE 3 – Courbes de survie

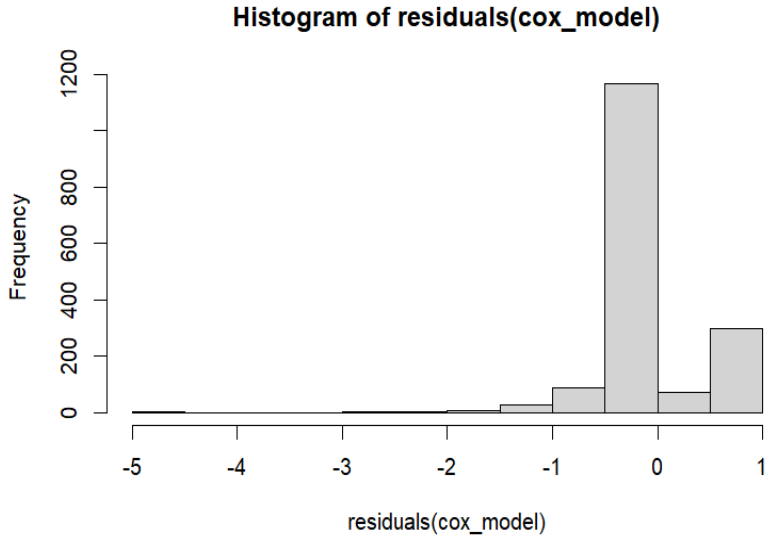

FIGURE 4 – Histogram des résidus

Références

[1] D. V. Dawson and B. L. Pihlstrom, “Application of biostatistics in dental public health.” on https://www.sciencedirect.com/topics/medicine-and-dentistry/proportional-hazards-model#:~:text=The%20Cox%20proportional%20hazards%20model%20is%20a%20frequently%20used%20approach,no%20distributional%20assumptions%20are%20required, 2021. Accessed : 2024-12-27.

[2] M. Marks and C. Roberts, “Poisson regression dataset for "learning clinical epidemiology with r" tutorial.” [Data Collection]. London School of Hygiene & Tropical Medicine, London, United Kingdom on https://datacompass.lshtm.ac.uk/id/eprint/609/, 2017.

[3] J. Rehm, D. Baliunas, G. L. G. Borges, K. Graham, H. Irving, T. Kehoe, C. D. Parry, J. Patra, S. Popova, V. Poznyak, M. Roerecke, R. Room, A. V. Samokhvalov, and B. Taylor, “The relation between different dimensions of alcohol consumption and burden of disease : an overview.” https://pubmed.ncbi.nlm.nih.gov/20331573/. Accessed : 2024-12-29.