Examen Modèle de durée

Consignes

— Rendu de cinq pages maximum (annexes non-incluses) par groupe de 5.

— Rendu pour le **17 Décembre**

— Nom du fichier attendu `duree_nom1_nom2_nom3.pdf`. Les documents rédigés en LaTeX, et un style d'écriture **épuré et lisible** seront favorisés.

— L'ensemble de l'analyse de données doit être réalisé avec R.

— L'ensemble du code commenté doit être mis en annexe. Les usages des LLM - type ChatGPT, Mystral - sont autorisés, mais uniquement dans un but de **soutien**. Toute portion de code générée à l'aide d'un LLM doit être signalée en commentaire dans l'annexe. Si le code est d'une complexité non-attendue, vous pouvez être convoqué pour expliquer votre code : si vous n'êtes pas en mesure de le faire, vous serez considéré comme ayant triché, et donc défaillant dans la matière.

— Tous les graphiques doivent comporter un titre, des axes clairement nommés, et indiquer explicitement la censure lorsqu'elle est présente.

Vous devez rédiger un rapport visant à démontrer l'intérêt des modèles de durée pour analyser l'un des phénomènes décrits dans la dernière section du présent document. Pour motiver votre démarche, vous vous appuierez sur **au moins trois articles scientifiques**. Une synthèse d'environ une demi-page suffit : il s'agit d'expliquer en quoi les modèles de durée sont adaptés au phénomène étudié et comment la littérature les utilise.

Votre analyse devra ensuite suivre les étapes suivantes :

— *(i)* préparer et mettre en forme la base de données, en définissant précisément la variable de durée et la structure de censure ;

— *(ii)* analyser les distributions des objets fondamentaux des modèles de durée (fonction de survie, fonction de hasard, taux de censure) ;

— *(iii)* estimer et commenter les courbes de Kaplan–Meier et de Nelson–Aalen ;

— *(iv)* discuter les covariables qui seraient pertinentes dans une analyse ultérieure, en expliquant leur rôle potentiel dans la dynamique de survie ou du risque ;

— *v)* conclure en expliquant vos résultats auprès d'un practicien ou un personnel de l'entreprise concerné n'ayant aucune connaissance en statistiques.

Important : vous ne devez pas estimer de modèle incluant des covariables (modèle de Cox, exponentiel, Weibull, ou tout autre modèle paramétrique ou semi-paramétrique). La réflexion sur les covariables doit rester théorique et prospective, en vue d'une extension future.

Structure attendue du rapport :

1. Introduction (motivation, littérature, question de recherche)
2. Données et construction des variables de durée
3. Analyse descriptive : distributions, courbes de survie
4. Estimation non paramétrique (Kaplan–Meier / Nelson–Aalen)
5. Discussion des covariables (sans estimation obligatoire)
6. Conclusion (incluant un paragraphe vulgarisé pour un public non technique)
7. Annexes

Bases de données disponibles pour le projet

Le choix du dataset conditionne la qualité de votre analyse ; vérifiez que les données contiennent bien une notion exploitable de durée. En cas de doute, contactez l'enseignant.

1. Employee Attrition (IBM HR Analytics)

Objet : durée dans l'entreprise et probabilité de départ volontaire.
Événement : Attrition (`Yes`) ; observation censurée sinon.

Liens :

— Page du projet : `https://github.com/ybifoundation/Dataset`

2. Telco Customer Churn (IBM Open Source)

Objet : durée d'abonnement à un service télécom.
Événement : résiliation (`Churn = Yes`).

Liens :

— Page du projet : `https://github.com/IBM/telco-customer-churn-on-icp4d`

3. Bike Sharing Dataset (UCI Machine Learning Repository)

Objet : données de location de vélos en libre-service, permettant de construire une variable de durée.
Événement : à définir (ex. : temps jusqu'au pic d'utilisation, jusqu'à une variation météo...).

Liens :

— Page du projet : `https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`

— Téléchargement direct (ZIP) : `https://archive.ics.uci.edu/ml/machine-learning-databases/00275/Bike-Sharing-Dataset.zip`

4. Kickstarter Projects (financement participatif)

Objet : durée d'une campagne avant succès ou échec.
Événement : financement atteint (succès).
Page du projet : `https://github.com/mkucz95/kickstarter_data?tab=readme-ov-file`

5. Heart Failure Clinical Records (UCI)

Objet : survie de patients ayant eu une insuffisance cardiaque.
Événement : décès pendant la période de suivi.

Page officielle du dataset :

`https://archive.ics.uci.edu/dataset/519/heart+failure+clinical+records`

Grille d'évaluation détaillée (100 points)

1. Structure et qualité rédactionnelle (20 points)

Excellent (16–20)	**Satisfaisant (10–15)**	**Insuffisant (<10)**
Rapport très bien structuré, rédaction claire, figures lisibles.	Structure correcte, quelques imprécisions.	Rapport désorganisé, nombreuses erreurs, figures absentes ou illisibles.

2. Construction des données et de la durée (20 points)

Définition précise de la durée et de la censure, préparation rigoureuse.	Définition correcte mais justification limitée.	Durée mal définie, censure ignorée, préparation incorrecte.

3. Analyse descriptive et objets des modèles de durée (20 points)

Analyse fine des distributions, graphiques lisibles et interprétés.	Analyse correcte mais superficielle.	Analyse insuffisante ou incorrecte, pas d'interprétation.

4. Estimation non paramétrique (Kaplan–Meier / Nelson–Aalen) (25 points)

Courbes correctement estimées, comparaison pertinente, interprétation solide.	Courbes présentes mais interprétation limitée.	Courbes absentes, erronées ou non interprétées.

5. Discussion des covariables (15 points)

Discussion rigoureuse des covariables pertinentes et de leur rôle potentiel.	Discussion partielle ou limitée.	Discussion absente ou incohérente.

6. Reproductibilité et qualité du code (10 points)

Code propre, commenté, intégralement reproductible.	Code fonctionnel mais peu commenté.	Code absent, inexploitable ou non conforme.

Attentes minimales pour valider le rapport

Pour être considéré comme complet et conforme, votre rapport doit impérativement satisfaire les exigences suivantes :

1. Construction rigoureuse de la variable de durée

- Définition explicite et justifiée de la durée analysée.
- Identification correcte de la censure et discussion du taux de censure.
- Code de préparation des données clair, reproductible et placé en annexe.

2. Production d'au moins quatre graphiques obligatoires

- Histogramme ou densité des durées.
- Courbe de Kaplan–Meier avec indication de la censure.
- Estimateur de Nelson–Aalen correctement présenté.
- Un graphique supplémentaire pertinent (ex. : distribution de la censure, etc.).

3. Interprétation technique et précise des résultats

- Analyse de la survie estimée (tendances, ruptures, comportement temporel).
- Discussion du risque (hazard) au cours du temps.
- Analyse critique de la qualité et des limites des données.

4. Mobilisation d'au moins trois articles scientifiques

- Synthèse d'une demi-page maximum.
- Mise en contexte méthodologique ou application des modèles de durée.
- Lien explicite avec votre problématique empirique.

5. Discussion approfondie des covariables

- Identification argumentée de 3 à 6 covariables pertinentes.
- Discussion de leur rôle potentiel sur la survie ou le hazard.
- Aucun modèle multivarié ne doit être estimé.

6. Conclusion vulgarisée

- Résumé clair et accessible à un praticien non statisticien.
- 10 lignes maximum.