On considère la durée d'abonnement chez un fournisseur de services de télécommunications (télévision, internet, téléphonie fixe et cellulaire). La base de données contient un échantillon aléatoire simple de $n = 1000$ observations comprenant les variables suivantes :

— `duree` : nombre de jours d'abonnement.
— `tchange` : nombre de jours d'abonnement au moment du changement de forfait.
— `actif` : variable binaire ; `0` si le client est inactif, `1` s'il est encore abonné.
— `nserv1` : nombre de services lors du contrat initial
— `nserv2` : nombre de services lors du changement de forfait le cas échéant (valeur manquante sinon.)
— `sexe` : variable indicatrice binaire ; `0` pour femmes, `1` pour hommes.

1. Combien y a-t-il d'observations censurées ?
2. Expliquez dans vos mots l'impact de la censure à droite (non-informative) sur l'inférence : qu'arrive-t-il si vous ignorez cette dernière et modélisez directement la durée d'abonnement sans ajustement aucun (par exemple, en calculant les statistiques descriptives de durée) ?
3. On ignore dans un premier temps l'impact du nombre de services sur la durée d'abonnement. Supposons que l'on veuille comparer les courbes de survie selon la tranche d'âge et le sexe en utilisant l'estimateur de Kaplan–Meier :
   — Testez l'égalité des courbes de survie à l'aide du test du log-rang. Écrivez les hypothèses nulle et alternative, la valeur de la statistique, la valeur-p et la conclusion du test.
   — Produisez un graphique des courbes de survie superposées pour chacun des sous-groupes avec intervalles de confiances ponctuels à 95%.
   — Rapportez l'estimé de la survie à 100 jours pour un homme de 25 ans qui est abonné à un seul service.
4. Le modèle de régression à risques proportionnels de Cox permettrait également de tester cette hypothèse. Quelles variables explicatives devriez-vous inclure pour ce faire dans le modèle et sous quelle forme ?
5. Expliquez les avantages et inconvénients liés à l'utilisation du modèle de Cox par rapport à la spécification de courbes de survie différentes pour chaque sous-groupe.
6. En supposant que l'effet du nombre de services est continu, ajustez un modèle à risque proportionnel en prenant en compte le fait que la valeur de la variable explicative `nserv` change dans le temps au temps `tchange`. Incluez également `age` (catégorie de référence (50, 90]) et `sexe` (sans interaction) comme variables explicatives. À l'aide de ce modèle :
   — Interprétez l'effet de la variable `sexe` sur le risque de désabonnement.
   — Toute chose étant égale par ailleurs, pour quelle catégorie d'âge la durée d'abonnement est-elle la plus élevée ? Justifiez votre réponse

Compilé le 05/04/2021 à 17:34