Université Paris Cité

BUT SD 3 FA EMS - Février 2025

TP - Analyse de données censurées

Modélisation statistique avancée

Réalisé par

Rachid SAHLI

Table des matières

1 Introduction

Ce TP s'inscrit dans le cadre du cours *Analyse de données censurées*, dispensé dans le module de *Modélisation statistique avancée* en troisième année de *BUT Science des Données* (parcours exploration et modélisation statistique). Son objectif est d'explorer les méthodes statistiques adaptées aux données censurées, en mettant en œuvre des techniques d'analyse et de modélisation spécifiques à ce type de données.

2 Présentation des données

On étudie la base de données churn qui contient des informations sur 7 032 clients d'une grande compagnie. Le but est d'étudier les facteurs de risque pour expliquer les résiliations des clients. La variable d'intérêt est donc le temps écoulé depuis la souscription d'un contrat dans la compagnie jusqu'à la résiliation du client. La variable tenure représente la durée observée et la variable Churn indique si ce temps est la durée d'intérêt (auquel cas Churn vaut "Yes") ou si ce temps est une censure (auquel cas Churn vaut "No"). Les covariables observées sont :

— Partner : "Yes" ou "No', indique si le client est partenaire de la compagnie,

— PhoneService : "Yes" ou "No', indique si le client a pris un abonnement téléphonique via la compagnie,

— InternetService : "No" si le client n'a pas pris d'abonnement internet via la compagnie ou "DSL" si le client a pas pris un abonnement DSL ou "Fiber optic" si le client a pas pris un abonnement avec la fibre,

— PaymentMethod :indique le modede paiementqui peut être"Bank transfer (automatic)", "Credit card (automatic)", "Electronic check" ou "Mailed check",

— Contract : indique le type de contrat souscrit qui peut être "Month-to-month", "One year" ou "Two year",

— PaperlessBilling : "Yes" ou "No', indique si le client reçoit des factures papiers,

— Dependents : "Yes" ou "No', indique si le client a des personnes à charge (par exemple s'il a des enfants).

3 Réponses aux questions

3.1 Importation des données et des librairies

```r
# Import library ----
library(survival) # Analyses de survie
library(lava) # Calcul et estimation de modèles stats
library(knitr) # Réalisation de tableau

# Import data ----
setwd("/Users/rs777/Documents/Projet-datascience/Statistique_survie/Programme/Tp2")
churn <- read.csv("Churn.csv")
```

3.2 Codage de la variable indicatrice de censure

Nous recodons la variable Churn, initialement au format caractère (“Yes”, “No”), en une variable indicatrice où :

- 0 représente l’occurrence de l’événement d’intérêt (churn),
- 1 indique une observation censurée.

Nous supprimons également les 11 observations où la variable tenure est égale à 0.

```r
churn$Churn <- ifelse(churn$Churn == "No", 0, 1) # Codage de la variable
churn <- churn[churn$tenure!=0,] # Suppression de ces valeurs
```

3.3 Courbe de survie estimé par Kaplan-Meier

L’estimateur de Kaplan-Meier est un estimateur non paramétrique de la fonction de survie $S(t)$, qui représente la probabilité qu’un individu survive au-delà d’un temps t.
Il est donné par la formule suivante :

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

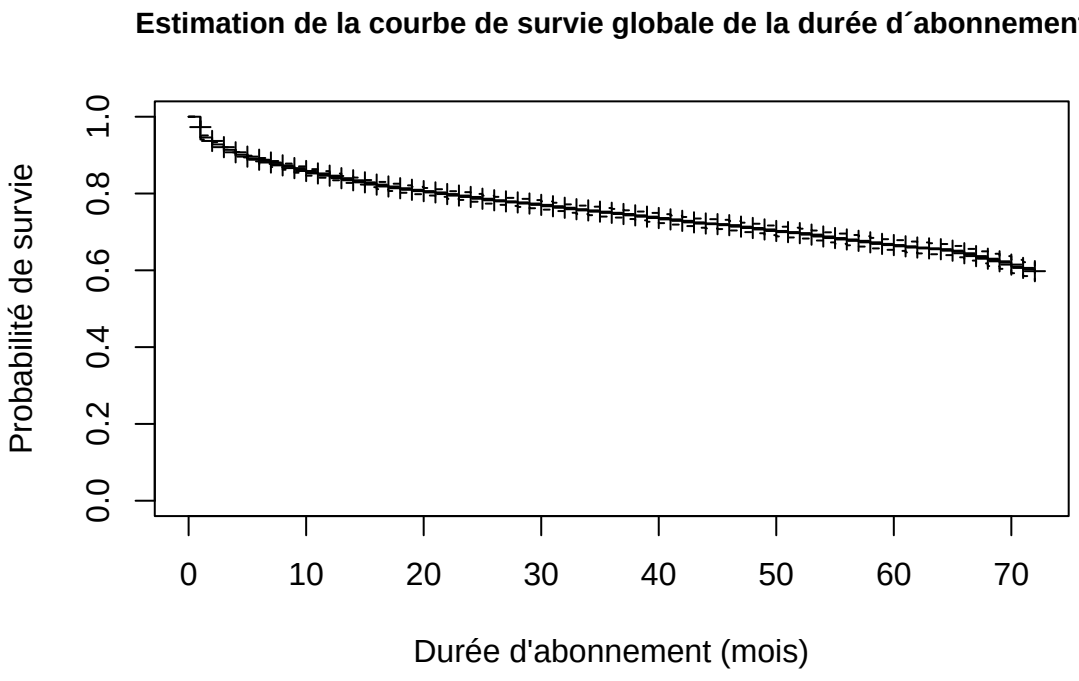
où :

- t_i sont les temps où un événement se produit,
- d_i est le nombre d'événements observés à t_i,
- n_i est le nombre d'individus encore en observation juste avant t_i.

L'estimateur est consistant et asymptotiquement normal sauf dans les "queues de distribution". De plus, on estime sa variance asymptotique par σ^2 par l'estimateur de Greenwood qui est un estimateur consistant (Greenwood, M. 1926 ; Breslow, N.E. et Crowley, J. J. 1974.).

On trace ci-dessous la courbe de survie globale de la durée d'abonnemenent estimé par la méthode de Kaplan-Meier.

```
KM <- survfit(Surv(churn$tenure, churn$Churn) ~ 1) # Estimation de la survie globale
# summary(KM, conf.type = "Plain") # Résultat de l'estimation
plot(KM,
     main = "Estimation de la courbe de survie globale de la durée d'abonnement",
     cex.main = 0.9, xlab = "Durée d'abonnement (mois)",
     ylab = "Probabilité de survie", conf.int = TRUE,
     mark.time = TRUE)
```



On observe une courbe décroissante. La probabilité de survie diminue au fur et à mesure que le temps passe. En effet, plus les individus restent abonnés pendant une période prolongée, plus ils

sont susceptibles de résilier leur abonnement.

Puis, nous estimons le premier quartile, qui correspond au temps pour lequel 25 % des individus ont déjà résilié leur abonnement.

```
quantile(KM, probs = 0.25) # Estimation du Q1
```

```
## $quantile
## 25
## 36
##
## $lower
## 25
## 32
##
## $upper
## 25
## 39
```

La probabilité que 25 % des individus aient résilié leur abonnement est de 36 mois (3 ans), avec un intervalle de 32 à 39 mois.

Enfin, nous estimons la probabilité d'avoir résilié son contrat dans la compagnie au bout de 4 ans (48 mois).

```
summary(KM, conf.type = "Plain", times = 48) # Estimation de la survie a 4 ans
```

```
## Call: survfit(formula = Surv(churn$tenure, churn$Churn) ~ 1)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    48   2303    1656    0.709 0.00632        0.697        0.721
```

À 4 ans, on estime que la probabilité de survie est de 0.709, soit environ 70.9 %. Cela signifie, qu'à 4 ans, environ 70.9% des individus restent abonnés à la compagnie, tandis que 29.1% ont résilié leur contrat. L'intervalle de confiance à 95 % indique que, si on répétait l'étude plusieurs fois, dans 95 % des cas, la probabilité que les clients restent abonnés pendant 4 ans serait entre 69,7 % et 72,1 %.

3.4 Analyse de la durée de souscription en fonction de variables qualitatives

Ici, nous allons tracer les estimateurs de Kaplan-Meier pour chaque modalité de nos variables qualitatives. Nous réaliserons ensuite le test du log-rang pour évaluer s'il existe des différences significatives entre les courbes de survie des différentes modalités. Ces analyses nous permettront d'identifier les variables qualitatives qui ont un impact sur la durée de souscription et de comparer les probabilités de survie en fonction des différentes catégories.

Le test du log-rang (Gehan, E. A. 1965 et Mantel, N. 1966) permet de tester la comparaison de deux courbes de survie. C'est un test non paramétrique asymptotique qui fonctionne en présence de données censurées.

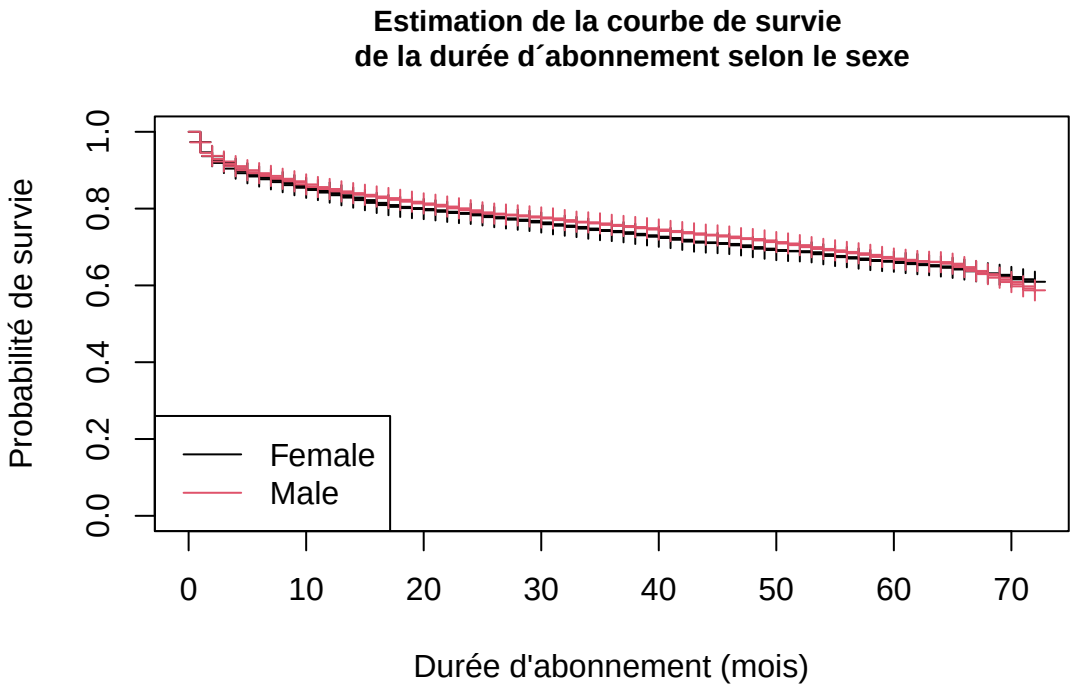
Notons S_A et S_B les fonctions de survie des deux groupes de survie A et B, ou souhaitera tester les hypothèses suivantes :

$$H_0 : S_A = S_B \quad \text{contre} \quad H_1 : S_A \neq S_B$$

3.4.1 Durée de souscription en fonction du genre

Nous débutons par estimer la durée de souscription en fonction du genre. Pour ce faire, nous utilisons la méthode de Kaplan-Meier pour estimer la courbe de survie selon cette variable (gender).

```
KM_gender <- survfit(Surv(churn$tenure, churn$Churn)
                     ~ churn$gender)# Estimation de la survie selon le sexe
plot(KM_gender,
     col = c(1,2),
     main = "Estimation de la courbe de survie
     de la durée d'abonnement selon le sexe",
     cex.main = 0.9,
     xlab = "Durée d'abonnement (mois)",
     ylab = "Probabilité de survie",
     mark.time = TRUE)
legend("bottomleft", legend = levels(as.factor(churn$gender)), lty=1, col=c(1,2))
```



Sur le graphique ci-dessus, les courbes de survie selon le genre sont très proches l'une de l'autre, se croisant à plusieurs reprises. Cela indique qu'il n'y a pas de différence nette et marquée dans la durée d'abonnement en fonction du sexe. De plus, les courbes se chevauchent souvenet. On compare ensuite les quantiles des deux distributions.

```
quantile(KM_gender, probs = 0.25) # Estimation des mesures de position
```

```
## $quantile
##                      25
## churn$gender=Female 34
## churn$gender=Male   39
##
## $lower
##                      25
## churn$gender=Female 30
## churn$gender=Male   33
##
## $upper
##                      25
## churn$gender=Female 38
```

```
## churn$gender=Male   44
```

Les résultats du premier quartile montrent que parmi les 25 % des clients ayant la plus courte durée d'abonnement, les hommes ont tendance à rester abonnés un peu plus longtemps que les femmes. En effet, le premier quartile des femmes se situe à 34 mois, ce qui signifie que 25 % des femmes résilient leur abonnement avant ce délai. Pour les hommes, ce chiffre est de 39 mois, ce qui indique que 25 % des hommes résilient leur abonnement avant 39 mois. Bien que les intervalles de confiance des deux groupes se chevauchent, ce qui suggère une certaine incertitude quant à la différence, on peut observer en moyenne que les hommes ont tendance à rester abonnés plus longtemps que les femmes.

On réalise ensuite le test du log-rang pour vérifier si cette différence est significative.

```
survdiff(Surv(churn$tenure, churn$Churn) ~ churn$gender) # Test du log-rang
```

```
## Call:
## survdiff(formula = Surv(churn$tenure, churn$Churn) ~ churn$gender)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## churn$gender=Female 3483      939      923     0.261     0.526
## churn$gender=Male   3549      930      946     0.255     0.526
##
##  Chisq= 0.5  on 1 degrees of freedom, p= 0.5
```

Pour les femmes, 939 événements (résiliations d'abonnement) ont été observés parmi 3483 individus, contre 923 événements attendus. Chez les hommes, 930 événements ont été observés parmi 3549 individus, avec un nombre attendu de 946 événements. Les différences entre les valeurs observées et attendues sont relativement faibles, ce qui indique que les deux groupes suivent des tendances similaires.

La statistique du test du log-rang est de 0.5, avec 1 degré de liberté, et la p-valeur associé est de 0.5. Puisque cette valeur est supérieure au seuil α de 0.05, nous ne rejetons pas l'hypothèse nulle (H_0). Autrement dit, il n'y a pas de différence statistiquement significative entre les courbes de survie des hommes et des femmes en ce qui concerne la durée d'abonnement.

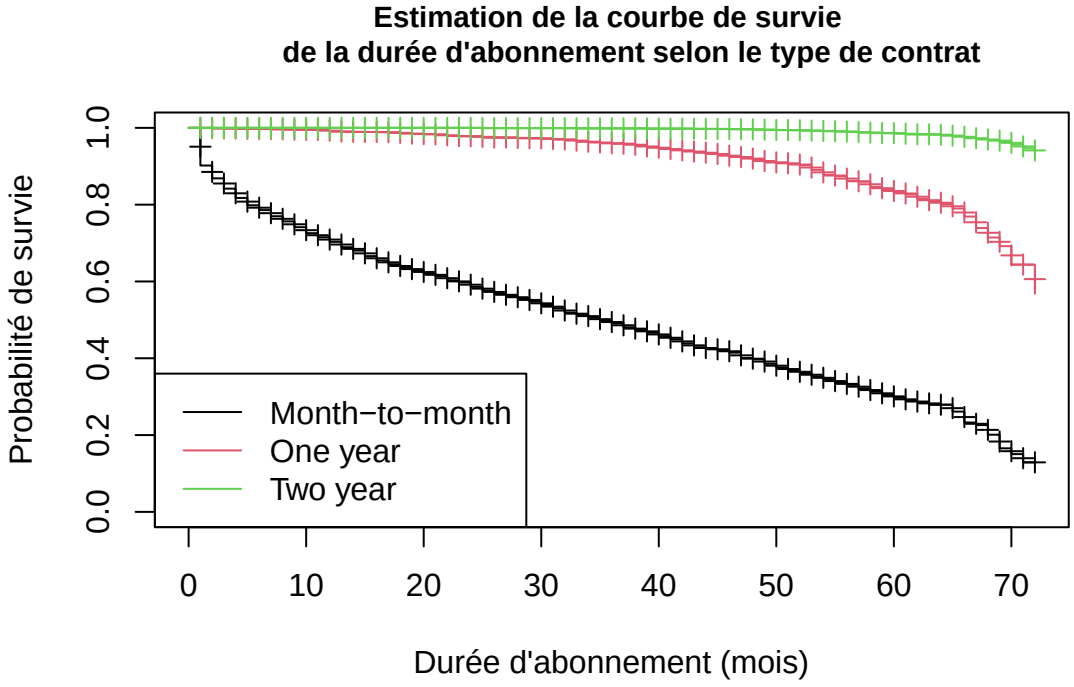
En résumé, ces résultats montrent que le genre n'a pas d'impact significatif sur la durée d'abonnement des clients. Cela confirme l'observation sur le graphique, où les courbes de survie des deux groupes

sont proches, ainsi que la faible différence au niveau du premier quartile entre les deux groupes.

3.4.2 Durée de souscription en fonction du contrat

Nous procédons ensuite à l'estimation de la durée de souscription en fonction du type de contrat. Il convient de rappeler que trois types de contrats sont proposés : "Month-to-month", "One year" et "Two year".

```
KM_contrat <- survfit(Surv(churn$tenure, churn$Churn)
                      ~ churn$Contract)
plot(KM_contrat,
     col = c(1,2,3),
     main = "Estimation de la courbe de survie
     de la durée d'abonnement selon le type de contrat",
     cex.main = 0.9,
     xlab = "Durée d'abonnement (mois)",
     ylab = "Probabilité de survie",
     mark.time = TRUE)
legend("bottomleft", legend = levels(as.factor(churn$Contract)), lty=1, col=c(1,2,3))
```



La courbe bleue représentant le contrat "Month-to-month" décroît beaucoup plus rapidement que

les deux autres. Cela indique que les clients ayant un contrat mensuel résilient leur abonnement plus tôt, avec une probabilité de survie qui diminue rapidement au fil du temps. La courbe verte, représentant le contrat "Two year", reste presque plate et très proche de 1. Les abonnées ayant choisi ce type de contrat de deux ans ont une probabilité de survie très élevée tout au long de la période. Ils sont donc beaucoup moins enclins à résilier leur abonnement. Enfin, la courbe rouge pour le contrat "One year" se situe entre les deux autres. Elle décroît plus lentement que celle du contrat mensuel, mais plus rapidement que celle du contrat de deux ans. Cela montre que les clients ayant souscrit à un contrat d'un an ont une probabilité de survie intermédiaire.

En somme, on observe sur ce graphique que la durée de souscription est fortement influencé par le type de contrat choisi.

On estime ci-dessous les quantiles de la courbe de survie pour chaque type de contrat afin d'analyser plus précisément la durée d'abonnement.

```
quantile(KM_contrat, probs = 0.25)
```

```
## $quantile
##                                 25
## churn$Contract=Month-to-month    9
## churn$Contract=One year         67
## churn$Contract=Two year         NA
##
## $lower
##                                 25
## churn$Contract=Month-to-month    8
## churn$Contract=One year         66
## churn$Contract=Two year         NA
##
## $upper
##                                 25
## churn$Contract=Month-to-month   10
## churn$Contract=One year         69
## churn$Contract=Two year         NA
```

Le type de contrat ne contient que des NA car les abonnés ayant un contrat de 2 ans (Two year) n'ont surement pas encore résilié leur abonnement au moment de l'observation, ce qui empêche le calcul des quantiles pour cette catégorie. Autrement dit, la majorité des utilisateurs ayant ce type de contrat sont encore en vie dans le processus de souscription, ce qui empêche l'estimation du quantile pour les 25 % de résiliations, car aucun événement de résiliation n'a eu lieu jusqu'à ce point.

Les clients ayant un engagement mensuel résilient très rapidement, tandis que ceux sous contrat d'un an restent abonnés bien plus longtemps. Pour le contrat Month-to-Month, qui correspond à la durée d'abonnement la plus courte, le premier quartile est estimé à 9 mois, avec un intervalle de confiance compris entre 8 et 10 mois. Cela signifie que 25 % des abonnés ayant ce type de contrat résilient leur abonnement avant 9 mois.

En comparaison, pour le contrat One Year, le premier quartile est estimé à 67 mois, avec un intervalle de confiance entre 66 et 69 mois. Cela indique que 25 % des abonnés ayant souscrit à un contrat d'un an résilient leur abonnement après environ 5 ans et demi. La différence avec le contrat mensuel est donc très marquée.

Enfin, l'absence de valeur pour le contrat Two Year (NA) suggère une fidélité encore plus forte. En effet, cela signifie qu'au moins 75 % des abonnés ayant choisi ce type de contrat n'ont pas encore résilié leur abonnement, rendant ainsi impossible l'estimation du premier quartile. Ce résultat confirme que plus l'engagement contractuel est long, plus la probabilité de rester abonné est élevée.

Nous effectuons à nouveau le test du log-rang afin de déterminer si cette différence est statistiquement significative.

```r
survdiff(Surv(churn$tenure, churn$Churn) ~ churn$Contract) # Test du log rank
```

```
## Call:
## survdiff(formula = Surv(churn$tenure, churn$Churn) ~ churn$Contract)
##
##                                    N Observed Expected (O-E)^2/E (O-E)^2/V
## churn$Contract=Month-to-month 3875     1655      708      1265      2304
## churn$Contract=One year       1472      166      471       197       270
## churn$Contract=Two year       1685       48      690       597      1061
##
##  Chisq= 2353  on 2 degrees of freedom, p= <2e-16
```

Pour les clients avec un contrat Month-to-Month, 1655 événements (résiliations) ont été observés sur 3875 abonnés, alors que le nombre attendu était de 708. Pour les clients avec un contrat One Year, 166 événements ont été observés sur 1472 abonnés, contre 471 attendus. Pour les clients avec un contrat Two Year, seulement 48 résiliations ont été observées sur 1685 abonnés, alors que 690 étaient attendues.

La statistique du test du log-rang est de 2353, avec 2 degrés de liberté, et une p-valeur inférieure à 2e-16. Cette p-valeur étant largement en dessous du seuil de significativité $\alpha = 0.05$, nous rejettons l'hypothèse nulle (H_0). Cela indique qu'au moins une des courbes de survie diffère significativement des autres en fonction du type de contrat.

En somme, nous en concluons que le type de contrat semble etre lié à la durée de souscription.

3.5 Modèle de Cox en fonction de la variable InternetService

Le modèle de Cox est une modèle de régression semi-paramétrique. Il permet d'analyser l'effet de plusieurs variables explicatives sur le temps avant la survenu d'un évènement. Il s'écrit de la manière suivante :

$$h(t|Z_{i1}, \ldots, Z_{ip}) = h_0(t) \exp(\theta_1 Z_{i1} + \cdots + \theta_p Z_{ip}) = h_0(t) \exp(\theta_0 Z_i)$$

Ce modèle fait les deux hypothèses suivantes sur les données :

- **Hypothèse des risques proportionnels :** Le rapport des risques instantanés ("hazard rate" en anglais) de deux patients est indépendant du temps.
- **Hypothèse de log-linéarité :** $\log(h(t|Z_{i1}, \ldots, Z_{ip})) = \log(h_0(t)) + \theta_0 Z_i$. Le logarithme du risque instantané est une fonction linéaire des Z_{ij}.

Nous construisons un modèle de Cox uniquement sur la variable InternetService qui dispose de trois modalités : DSL, Fiber optic, No. L'objectif est d'analyser l'impact de la variable InternetService sur la durée d'abonnement. Il va donc nous permettre de comparer les risques de résiliation de l'abonnement entre ces trois groupes, en prenant en compte les effets spécifiques de chaque groupe.

```
churn[sapply(churn, is.character)] <- lapply(churn[sapply(churn, is.character)], as.factor) #
churn$InternetService <- relevel(factor(churn$InternetService)
                                 , ref = "No") # 1 ere modalité par défaut
```

```r
fit_internet = coxph(Surv(churn$tenure, churn$Churn) ~ factor(churn$InternetService)) # Modèle
summary(fit_internet)
```

```
## Call:
## coxph(formula = Surv(churn$tenure, churn$Churn) ~ factor(churn$InternetService))
##
##   n= 7032, number of events= 1869
##
##                                          coef exp(coef) se(coef)      z
## factor(churn$InternetService)DSL         0.89233   2.44081  0.10502  8.497
## factor(churn$InternetService)Fiber optic 1.68678   5.40203  0.09809 17.196
##                                          Pr(>|z|)
## factor(churn$InternetService)DSL           <2e-16 ***
## factor(churn$InternetService)Fiber optic   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                          exp(coef) exp(-coef) lower .95
## factor(churn$InternetService)DSL             2.441     0.4097     1.987
## factor(churn$InternetService)Fiber optic     5.402     0.1851     4.457
##                                          upper .95
## factor(churn$InternetService)DSL             2.999
## factor(churn$InternetService)Fiber optic     6.547
##
## Concordance= 0.626  (se = 0.006 )
## Likelihood ratio test= 561.4  on 2 df,   p=<2e-16
## Wald test            = 446.2  on 2 df,   p=<2e-16
## Score (logrank) test = 516.7  on 2 df,   p=<2e-16
```

Les résultats montrent que les abonnés ayant un service Fiber optic ont un risque de résiliation bien plus élevé (environ 5.4 fois) par rapport à ceux sans service Internet (“No”). Les abonnés avec un service DSL ont également un risque de résiliation plus élevé (environ 2.44 fois) par rapport à ceux sans service Internet. Ces résultats sont confirmés par des p-values des test qui sont très faibles, ce

qui indique que ces différences sont significatives d'un point de vue statistique. Il y a bien un effet significatif sur la durée de souscription.

3.6 Modèle de Cox en fonction des variables InternetService et Contract

Nous rajoutons au modèle précédent la variable Contract.

```
fit_2 = coxph(Surv(churn$tenure, churn$Churn) ~ churn$InternetService + churn$Contract) # Modè
summary(fit_2)
```

```
## Call:
## coxph(formula = Surv(churn$tenure, churn$Churn) ~ churn$InternetService +
##     churn$Contract)
##
##   n= 7032, number of events= 1869
##
##                                    coef exp(coef) se(coef)       z Pr(>|z|)
## churn$InternetServiceDSL          0.29090   1.33763  0.10567   2.753  0.00591
## churn$InternetServiceFiber optic  0.58805   1.80048  0.10002   5.879 4.12e-09
## churn$ContractOne year           -2.10046   0.12240  0.08425 -24.932  < 2e-16
## churn$ContractTwo year           -4.03059   0.01776  0.15799 -25.512  < 2e-16
##
## churn$InternetServiceDSL         **
## churn$InternetServiceFiber optic ***
## churn$ContractOne year           ***
## churn$ContractTwo year           ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                                  exp(coef) exp(-coef) lower .95 upper .95
## churn$InternetServiceDSL           1.33763     0.7476   1.08739   1.64546
## churn$InternetServiceFiber optic   1.80048     0.5554   1.47995   2.19043
## churn$ContractOne year             0.12240     8.1699   0.10377   0.14438
## churn$ContractTwo year             0.01776    56.2940   0.01303   0.02421
```

```
##
## Concordance= 0.785  (se = 0.004 )
## Likelihood ratio test= 2678  on 4 df,   p=<2e-16
## Wald test            = 1326  on 4 df,   p=<2e-16
## Score (logrank) test = 2410  on 4 df,   p=<2e-16
```

Dans ce modèle, le rapport de risque pour les abonnés à la DSL est de 1,34, ce qui signifie que le risque de résiliation est 1,34 fois plus élevé pour un abonné ayant la DSL par rapport à un abonné sans abonnement Internet. Pour les abonnés à la fibre optique, ce rapport de risque est de 1,8, ce qui indique un risque de résiliation 1,8 fois plus élevé par rapport aux abonnés sans abonnement Internet. Les p-valeurs associées aux deux modalités de la variable InternetService sont extrêmement faibles (inférieures à 0,001), ce qui rend ces résultats statistiquement très significatifs.

Concernant la variable Contract, pour la modalité “One year”, le rapport de risque est de 0,122, ce qui indique que les abonnés ayant un contrat d’un an ont un risque de résiliation considérablement plus faible que ceux ayant un contrat “Month-to-month”, avec une réduction de 87,76% du risque. En revanche, pour les abonnés ayant un contrat de deux ans, le rapport de risque est de 0,0178, suggérant une réduction encore plus marquée du risque de résiliation, soit une diminution de 98,22% par rapport aux abonnés “Month-to-month”.

Les rapports de risque subissent une modification importante lorsqu’on ajoute la variable Contract dans le modèle. En effet, sans cette variable, les abonnés ayant la DSL ou la fibre optique présentent un risque de résiliation beaucoup plus élevé par rapport aux abonnés sans abonnement Internet. Cependant, une fois que l’on prend en compte le type de contrat, les rapports de risque changent de manière significative. Les abonnés ayant un contrat “Month-to-month” sont exposés à un risque de résiliation nettement plus élevé que ceux ayant un contrat d’un an ou de deux ans, ce qui montre l’importance de prendre en considération cette variable pour mieux comprendre les facteurs influençant la résiliation des abonnements.

On observe le lien entre ces deux variables.

```
kable(prop.table(table(churn$Contract,churn$InternetService),margin=1), caption = "Tableau de
```

Table 1: Tableau de contingence entre Contract et InternetService

	No	DSL	Fiber optic
Month-to-month	0.1352258	0.3156129	0.5491613
One year	0.2466033	0.3872283	0.3661685
Two year	0.3756677	0.3697329	0.2545994

Les abonnés "Month-to-month" ont une plus grande proportion d'abonnés à la fibre optique, tandis que ceux ayant un abonnement de "Two year" ont une proportion plus élevée d'abonnés sans abonnement Internet et un pourcentage plus équilibré entre DSL et fibre optique.

Nous réalisons un test du chi-carré afin de vérifier l'existence d'une association significative entre les variables Contract et InternetService.

Les hypothèses du test sont les suivantes :

H_0 : Contract $\perp$ InternetService (indépendants) contre H_1 : Contract $\not\perp$ InternetService (sont liés)

```
chisq.test(table(churn$Contract,churn$InternetService))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(churn$Contract, churn$InternetService)
## X-squared = 597.07, df = 4, p-value < 2.2e-16
```

Les résultats montre qu'il existe une association statistiquement significative entre la variable Contract et InternetService. La p-valeur est nettement inférieur au seuil $\alpha = 0.05$. On rejette donc l'hypothèse H_0 et nous en concluons que les deux distributions ne sont pas indépendantes.

3.7 Modèle de Cox avec la variable Dependents

Nous commencerons par construire un modèle de Cox en utilisant uniquement la variable Dependents. Ensuite, nous élargirons progressivement ce modèle en ajoutant d'autres variables.

```
fit_3 = coxph(Surv(churn$tenure, churn$Churn) ~ factor(churn$Dependents))
summary(fit_3)
```

```
## Call:
## coxph(formula = Surv(churn$tenure, churn$Churn) ~ factor(churn$Dependents))
##
##   n= 7032, number of events= 1869
##
##                                  coef exp(coef) se(coef)      z Pr(>|z|)
## factor(churn$Dependents)Yes -0.89981   0.40665  0.06105 -14.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                              exp(coef) exp(-coef) lower .95 upper .95
## factor(churn$Dependents)Yes    0.4066      2.459    0.3608    0.4583
##
## Concordance= 0.585  (se = 0.005 )
## Likelihood ratio test= 259.8  on 1 df,   p=<2e-16
## Wald test            = 217.2  on 1 df,   p=<2e-16
## Score (logrank) test = 232.2  on 1 df,   p=<2e-16
```

3.8 Modèle de Cox complet

3.9 Interaction entre Partner et InternetService

4 Conclusion