TD3 – Odds ratio et linéarité du logit

Partie 3 – Significativité globale du modèle, calcul et interprétation des OR

1. **Modélisation de `survived` en fonction de l'intercept uniquement**
   (a) Donner l'odds associé à ce modèle et le commenter.

   ```
   surv.null <- glm(survived ~ 1, family = binomial, data = df)
   exp(coef(surv.null))
   ```

2. **Modélisation de `survived` en fonction du sexe du passager**
   (a) Réaliser un test du rapport de vraisemblance pour cette variable explicative. Poser les hypothèses et conclure.

   ```
   1-pchisq(surv.sex$null.deviance-surv.sex$deviance,
            surv.sex$df.null-surv.sex$df.residual)
   ```

   (b) Calculer l'OR en prenant comme modalité de référence `sex = 'male'`.
   (c) Interpréter l'OR : quelle est l'impact sur la probabilité de survie lorsque l'on est une femme, par rapport aux hommes ?
3. **Modélisation de `survived` en fonction de la classe du passager**
   (a) Réaliser un test du rapport de vraisemblance pour cette variable explicative. Poser les hypothèses et conclure.
   (b) Calculer les OR en considérant la modalité de référence `pclass = '3'`.
   (c) Interpréter les OR
   (d) Interpréter l'OR associé aux modalités `pclass = '1'` et `pclass = '2'` en terme d'augmentation des chances de survie.
4. **Modélisation de `survived` en fonction de l'âge**
   (a) Calculer l'OR associé à la variable `age`
   (b) Interpréter l'OR : quel est l'impact sur la probabilité de survie lorsque l'âge augmente de 1 année ? de 10 ans ?
   (c) Le modèle est-il globalement significatif ? Poser les hypothèses et conclure.
   (d) Le test précédent vous fait-il revoir votre interprétation des OR ?
   (e) Interpréter le resultat des deux commandes suivantes :

   ```
   confint(surv.age)
   exp(confint(surv.age))
   ```

Partie 4 – Linéarité du logit

La régression logistique fait l'hypothèse de linéarité du logit. Cette hypothèse n'a pas d'impact dans le cas de variables explicatives binaires mais doit être vérifiée dans le cas de variables explicatives continues. On voit bien notamment que l'OR est constant quelle que soit la valeur de x_i. Cette hypothèse n'est pas forcément vérifiée : par exemple, si on souhaite modéliser la probabilité d'avoir la maladie d'Alzheimer, l'augmentation des chances est bien plus importante entre `60 et 70 ans` qu'entre `20 et 30 ans`. On doit donc vérifier l'hypothèse de linéarité du logit avant d'introduire une variable quantitative dans le modèle. Dans le cas où l'hypothèse n'est pas vérifiée, on peut alors rechercher une transformation de la variable qui la rendrait linéaire, soit découper la variable en classes.
Pour vérifier l'hypothèse, on commence par découper la variable en classes

```
bornes <- c(quantile(df$age, probs = seq(0, 1, by = 0.2)))
df$age_discret <- cut(df$age, breaks=bornes)
```

et on récupère la probabilité d'être $y = 1$ dans chaque classe.

```
pix = table(df$age_discret, df$survived)/rowSums(table(df$age_discret, df$survived))
```

On peut ensuite en déduire le logit associé à chaque classe :

```
logit=log((pix[,2])/(1-pix[,2]))
```

que l'on trace en fonction des centres de classes

```
x = (c(0,19,25,31,42)+c(19,25,31,42,80))/2
plot(x, logit)
lines(x, lm(logit~x)$fit)
```

Si les points forment (à peu près) une droite, on accepte l'hypothèse de linéarité du logit et on peut introduire la variable explicative continue dans le modèle.
Que concluez-vous ici quand à l'hypothèse de linéarité du logit pour la variable `age` ?