Régression logistique 2019–2020

TD2 – Mise en oeuvre de la régression logistique sous R

**Les données** Le naufrage du Titanic est sûrement le plus célèbre des accidents maritimes de l'époque contemporaine. Le 15 avril 1922, il entraîna la mort de 1502 personnes sur les 2224 passagers présents à bord dont le pauvre Léonardo, qui n'en demandait pas tant. Bien que des tentatives d'explication aient été abordées dans le film éponyme d'une longueur presque crapuleuse, les causes qui ont entraîné la survie ou la mort des passagers demeurent en grande partie inexpliquées. L'objectif sera donc pour vous de démêler les fils de la fatalité à bord du Titanic. Pour cela nous allons nous intéresser à la modélisation de la survie des passagers en fonction de données explicatives recueillies sur un échantillon de 1310 personnes présentes sur le navire. Les données sont les suivantes :

- `pclass` : la classe de la cabine réservée par le passager (1ère, 2ème ou 3ème classe)
- `survived` : si le passager a survécu (`survived`=1) ou non (`survived`=0)
- `name` : le nom du passager
- `sex` : le sexe du passager (`male` ou `female`)
- `age` : l'âge du passager
- `sibsp` : le nombre de relations familiales au même niveau (frère/soeur ou mari/femme) du passager présents sur le Titanic
- `parch` : le nombre de relations familiales d'un niveau diérent (parents ou enfants) du passager présents sur le Titanic
- `embarked` : le port d'embarquement du passager (`C` = Cherbourg, `Q` = Queenstown, `S` = Southampton)

On cherche à prédire la survie `survived`= 1 en fontion de toutes les variables.

## Partie 1 – Lecture du fichier et préparation des données

1. Lire le jeu de données et répérer ses variables

```
df <- read.table("titanic.csv", sep=";", header=TRUE, dec=",", na.strings = "")
```

2. Identifier les variables qualitatives et changer leur type en `factor`

```
class(df$pclass)
df$pclass <- factor(df$pclass)
str(df)
```

3. Certains individus ont des données manquantes (l'âge notamment). Supprimer ces individus du fichier. Combien d'individus comporte finalement le jeu de données avec lequel on va travailler ?

```
df <- df[complete.cases(df),]
```

**Partie 2 – Prédiction de `survived` en fonction d'une seule variable**

1. Que modélise t-on ? Indiquer à R la modalité à prédire.

   ```
   df$survived <- relevel(df$survived, ref="0")
   ```

2. Combien de passagers ont-ils survécu au naufrage ?

   ```
   table(df$survived)
   ```

3. **Modélisation de `survived` en fonction de l'intercept uniquement**
   (a) Écrire le modèle. Dans un premier temps, en donner la formulation générale, puis indiquer spécifiquement le nom des variables, ainsi que l'estimation ponctuelle des coefficients dans l'équation.

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

   (b) Retrouver à la main la valeur du coefficient $b_0$ donné par R.
   (c) Donner le modèle que l'on aura obtenu si on avait modélisé `survived`$= 0$.
4. **Modélisation de `survived` en fonction de l'âge**
   (a) Écrire le modèle. Dans un premier temps, en donner la formulation générale, puis indiquer spécifiquement le nom des variables, ainsi que l'estimation ponctuelle des coefficients dans l'équation.

   ```
   surv.age <- glm(survived ~ age, family = binomial, data = df)
   summary(surv.age)
   ```

   (b) Qu'est ce qui change si, au lieu de modéliser le faible poids de naissance, on choisit de modéliser le non faible poids de naissance ?
   (c) Quelle est la probabilité pour un individu de 20 ans au moment du drame, tel que Jack, de survivre au naufrage du Titanic ? un individu de 28 ans, telle que Rose, de survivre au naufrage du Titanic ?
   (d) Soumettre le code suivant et expliquer la sortie.

   ```
   newdata = data.frame(age = 20)
   lin = predict(surv.age, newdata)
   exp(lin)/(1+exp(lin))
   ```

5. **Modélisation de `survived` en fonction du sexe du passager**
   (a) Écrire le modèle en considérant la modalité de référence `sex = 'male'`. Dans un premier temps, en donner la formulation générale, puis indiquer spécifiquement le nom des variables, ainsi que l'estimation ponctuelle des coefficients dans l'équation.

   ```
   df$sex<- relevel(df$sex, ref='male')
   ```

   (b) Écrire le modèle en considérant la modalité de référence `sex = 'female'`. Dans un premier temps, en donner la formulation générale, puis indiquer spécifiquement le nom des variables, ainsi que l'estimation ponctuelle des coefficients dans l'équation.
   (c) Montrer que les 2 modèles sont bien équivalents.
   (d) Calculer la probabilité de survie pour la modalité `sex = 'female'` pour les deux modèles.

(e) Retrouver à la main la valeur de $b_0$ lorsque la modalité de référence est `sex = 'female'` et `sex = 'male'`.

```
table(df$survived, df$sex)
```

6. **Modélisation de `survived` en fonction de la classe du passager**

(a) Écrire le modèle en considérant la modalité de référence `pclass = '1'`. Dans un premier temps, en donner la formulation générale, puis indiquer spécifiquement le nom des variables, ainsi que l'estimation ponctuelle des coefficients dans l'équation.

(b) Écrire le modèle en considérant la modalité de référence `pclass = '2'`. Dans un premier temps, en donner la formulation générale, puis indiquer spécifiquement le nom des variables, ainsi que l'estimation ponctuelle des coefficients dans l'équation.

(c) Écrire le modèle en considérant la modalité de référence `pclass = '3'`. Dans un premier temps, en donner la formulation générale, puis indiquer spécifiquement le nom des variables, ainsi que l'estimation ponctuelle des coefficients dans l'équation.

(d) Montrer que les 3 modèles sont équivalents.

(e) Calculer la probabilité d'être `survived=1` pour les 3 valeurs de `pclass`. Commenter.