Modèles de comptage avec R

Tuto@Mate, 21 mai 2024

Joseph Larmarange

Variable de type comptage

- *outcome* correspondant à un nombre entier positif
- souvent, nombre d'occurrences d'un évènement
- modèles linéaires et logistiques non adaptés

Modèle de Poisson

Un premier exemple

Descendance atteinte par des femmes à l'âge de 30 ans

- jeu de données `fecondite` fourni par le package `{questionr}`
- contient 3 tables : `menages`, `femmes` et `enfants`

Aperçu des données

```r
library(tidyverse)
library(labelled)
data("fecondite", package = "questionr")
enfants |> look_for()
```

```
 pos variable       label                        col_type missing values
 1   id_enfant      Identifiant de l'enfant      dbl      0
 2   id_femme       Identifiant de la mère       dbl      0
 3   date_naissance Date de naissance            date     0
 4   sexe           Sexe de l'enfant             dbl+lbl  0       [1] masculin
                                                                  [2] féminin
 5   survie         L'enfant est-il toujours en~ dbl+lbl  0       [0] non
                                                                  [1] oui
 6   age_deces      Age au décès (en mois)       dbl      1442
```

Aperçu des données

```
library(tidyverse)
library(labelled)
data("fecondite", package = "questionr")
enfants |> look_for()
```

```
 pos variable       label                        col_type missing values
 1   id_enfant      Identifiant de l'enfant      dbl      0
 2   id_femme       Identifiant de la mère       dbl      0
 3   date_naissance Date de naissance            date     0
 4   sexe           Sexe de l'enfant             dbl+lbl  0       [1] masculin
                                                                  [2] féminin
 5   survie         L'enfant est-il toujours en~ dbl+lbl  0       [0] non
                                                                  [1] oui
 6   age_deces      Age au décès (en mois)       dbl      1442
```

Les données sont labellisées –> conversion en facteurs avec `labelled::unlabelled()`

```
femmes <-
  femmes |>
  unlabelled()
enfants <-
  enfants |>
  unlabelled()
```

Préparation des données

Calcul de l'âge exact des mères à la naissance avec `lubridate::time_length()`

```
enfants <-
  enfants |>
  left_join(
    femmes |>
      select(id_femme, date_naissance_mere = date_naissance),
    by = "id_femme"
  ) |>
  mutate(
    age_mere = time_length(
      date_naissance_mere %--% date_naissance,
      unit = "years"
    )
  )
```

Préparation des données

Calcul de l'âge exact des mères à la naissance avec `lubridate::time_length()`

```
enfants <-
  enfants |>
  left_join(
    femmes |>
      select(id_femme, date_naissance_mere = date_naissance),
    by = "id_femme"
  ) |>
  mutate(
    age_mere = time_length(
      date_naissance_mere %--% date_naissance,
      unit = "years"
    )
  )
```

Comptons, par femme, le nombre d'enfants nés avant l'âge de 30 ans

```
femmes <-
  femmes |>
  left_join(
    enfants |>
      filter(age_mere < 30) |>
      group_by(id_femme) |>
      count(name = "enfants_avt_30"),
    by = "id_femme"
  ) |>
  tidyr::replace_na(list(enfants_avt_30 = 0L))
```

Préparation des données (2)

Calcul de l'âge des femmes au moment de l'enquête et recodage du niveau d'éducation

```
femmes <-
  femmes |>
  mutate(
    age = time_length(
      date_naissance %--% date_entretien,
      unit = "years"
    ),
    educ2 = educ |>
      fct_recode(
        "secondaire/supérieur" = "secondaire",
        "secondaire/supérieur" = "supérieur"
      )
  )
```

Préparation des données (2)

Calcul de l'âge des femmes au moment de l'enquête et recodage du niveau d'éducation

```
femmes <-
  femmes |>
  mutate(
    age = time_length(
      date_naissance %--% date_entretien,
      unit = "years"
    ),
    educ2 = educ |>
      fct_recode(
        "secondaire/supérieur" = "secondaire",
        "secondaire/supérieur" = "supérieur"
      )
  )
```

Enfin, nous n'allons garder que les femmes âgées d'au moins 30 ans au moment de l'enquête.

```
femmes30p <-
  femmes |>
  filter(age >= 30)
```

Calcul du modèle de Poisson

- fonction `stats::glm()` en précisant `family = poisson`
- réduction par minimisation de l'AIC avec `stats::step()`
- fonction de lien logarithmique (*log*) –> exponentielle des coefficients s'interprète comme un risque relatif

```
mod1_poisson <- glm(
  enfants_avt_30 ~ educ2 + milieu + region,
  family = poisson,
  data = femmes30p
)
mod1_poisson <- step(mod1_poisson)
```

```
Start:  AIC=1013.81
enfants_avt_30 ~ educ2 + milieu + region

         Df Deviance    AIC
- region  3   686.46 1010.6
<none>        683.62 1013.8
- milieu  1   686.84 1015.0
- educ2   2   691.10 1017.3

Step:  AIC=1010.65
enfants_avt_30 ~ educ2 + milieu

         Df Deviance    AIC
<none>        686.46 1010.6
- milieu  1   691.30 1013.5
- educ2   2   693.94 1014.1
```

Tableau des coefficients

```
library(gtsummary)
theme_gtsummary_language("fr",
  decimal.mark = ",", big.mark = " "
)
mod1_poisson |>
  tbl_regression(exponentiate = TRUE) |>
  bold_labels()
```

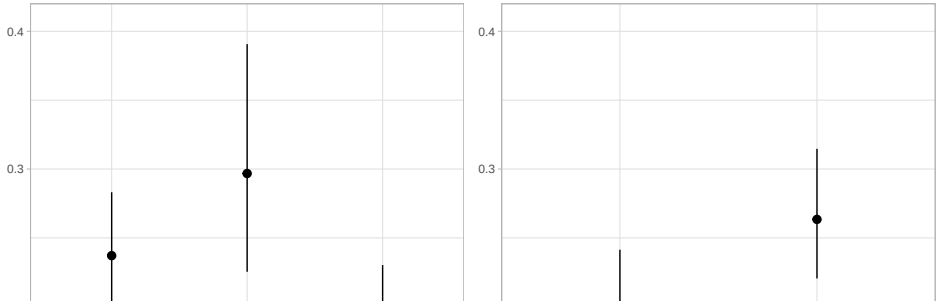
Caractéristique	**IRR**	**95% IC**	**p-valeur**
Niveau d'éducation			
aucun	—	—	
primaire	1,25	0,90 – 1,72	0,2
secondaire/supérieur	0,53	0,27 – 0,96	0,052
Milieu de résidence			
urbain	—	—	
rural	1,42	1,04 – 1,98	0,032

```
library(ggstats)
mod1_poisson |>
  ggcoef_table(exponentiate = TRUE)
```

Interprétation des coefficients

- Le modèle de Poisson modélise le nombre moyen d'évènements.
- Le RR pour la modalité *secondaire/supérieur* est de 0,5 : indépendamment des autres variables du modèle, la descendance atteinte moyenne de ces femmes est moitié moindre que celle des femmes de la modalité de référence.
- Vérification visuelle avec un graphique des prédictions marginales moyennes.

```
mod1_poisson |>
  broom.helpers::plot_marginal_predictions() |>
  patchwork::wrap_plots() &
  ggplot2::scale_y_continuous(limits = c(0, .4))
```



Évaluation de la surdiespersion