Statistics in Health Sciences

Assignment 2

Modeling incidence rates with generalized linear models

Contents

Instructions

1. Respect the numbering sections.
2. For all questions (if not stated otherwise):
    - Write the question
    - Answer the question
    - Show your R code (i.e. set `echo = TRUE`)
    - Show your R outputs
3. All inline results must be included using `\Sexpr`.
4. All tables and figures must be self-contained. I.e. captions must give all details needed to fully understand the contents.
5. Output shown in tables must be created using `xtable` or `kable`.
6. Output shown in figures must be created using an R chunk. The appearance of the figures can be controlled with the chunk options. For instance, width and height can be controlled with `fig.width` and `fig.height`, respectively (by default, in inches); the figure width relative to the page width can be controlled with `out.width`. Further details can be found at `https://yihui.org/knitr/options/#plots`

1 Introduction

In 1961, Doll and Hill[1] sent out a questionnaire to all men on the British Medical Register asking about their smoking habits. Almost 70% of such men replied. Death certificates were obtained for medical practitioners and causes of death were assigned on the basis of these certificates. The `breslow` data set[2] (attached in the `breslow.RData` file) contains the person-years of observations and deaths from coronary artery disease accumulated during the first ten years of the study. Such data are shown in Table 1.

	Person-years		Coronary deaths	
Age	Nonsmokers	Smokers	Nonsmokers	Smokers
35-44	18790	52407	2	32
45-54	10673	43248	12	104
55-64	5710	28612	28	206
65-74	2585	12663	28	186
75-84	1462	5317	31	102

Table 1: Data on coronary death rates.

The aim of this exercise is to analyze the relationship between incidence of coronary deaths and both smoke status (exposure of interest) and age (as a potential confounder).

2 Comparing incidence rates

1. Add extra columns in Table 1 for:

   (a) $I_{r_{ij}}$, the sample coronary death rates per 1000 person-years for smoke status i ($i = 0$ for nonsmokers and $i = 1$ for smokers) and age group j ($j = 0, 1, \ldots, 4$ for 35-44, 45-54, ..., 75-84, respectively) (two columns).

   (b) $IRR_j = \frac{I_{r_{1j}}}{I_{r_{0j}}}$, the incidence rate ratio for smokers vs. nonsmokers for the age group j (one column).

   Print the updated table as Table 2, with a proper caption.

   Show your code here for creating Table 2.

2. For age group 45-54, complete the following sentences with numbers and proper units **(using `\Sexpr` in your `Rnw` document)**:

   (a) The incidence rate among smokers was ...

   (b) The incidence rate among nonsmokers was ...

   (c) The incidence rate ratio was ...

   (d) The p-value of the Wald test to decide if the incidence rate among smokers is the same than among nonsmokers was ...

   (e) Write a paragraph for the interpretation of previous results (including all quantities).

3. For age group 75-84, complete the following sentences with numbers and proper units **(using `\Sexpr` in your `Rnw` document)**:

(a) The incidence rate among smokers was ...

(b) The incidence rate among nonsmokers was ...

(c) The incidence rate ratio was ...

(d) The p-value of the Wald test to decide if the incidence rate among smokers is the same than among nonsmokers was ...

(e) Write a paragraph for the interpretation of previous results (including all quantities).

4. Create a figure, named Figure 1, as follows:

   (a) Figure 1 must include two plots, one of them on the left and another on the right. **Hint:** Use `par(mfrow = c(1, 2), ...)`.

   (b) Both plots must represent age groups in the horizontal axis and rate ratios in the vertical axis.

   (c) The plot on the right must represent the rate ratios in logarithmic scale. **Hint:** Use `plot(log = "y", ...)`.

   (d) Both plots must include proper labels in both axes and proper ticks labels (for instance, horizontal axis must show labels 35-44, 45-54, ...). **Hint:** Use `plot(xaxt = "n", ...)` and then use `axis(1, at = ..., labels = ...)`.

   (e) Figure 1 must include a detailed caption.

   Show your R code here for creating Figure 1.

5. According to Figure 1:

   (a) Is the rate ratio greater than 1 for all age groups? What does it means?

   (b) Is the rate ratio constant over age groups? What does it means?

3 Modeling incidence rates with a Generalized Linear Model

Suppose we are interested in modeling the coronary deaths rate, $I = \frac{I}{\Delta t}$, where I is the coronary deaths count and Δt is follow-up (in person-years), as a function of smoking status (E), which is the exposure of interest, and age group (A). We suspect that smoking status could interact with age in the effect on coronary mortality. Hence, we consider the Generalized Linear Model (GLM)[1] specified in equations (1) to model the coronary deaths rate for individuals with smoking status i and in age group j, $I_{r_{ij}}$:

[1] You can find the LaTeX code to write model (1) in the file `model.tex`.

$$
\begin{cases}
\bullet \text{ Probability distribution:} \\
\qquad I_{r_{ij}} = \frac{I_{ij}}{\Delta t_{ij}}, \qquad I_{ij} \sim \text{Pois}(\lambda_{ij}), \qquad \Delta t_{ij} \text{ is the follow-up;} \\
\bullet \text{ Model for the mean:} \\
\qquad \log\left(\frac{\lambda_{ij}}{\Delta t_{ij}}\right) = L(E_i, A_j) \text{ (the linear predictor)} \\
\qquad \begin{aligned} L(E_i, A_j) &= \alpha + \beta_i E_i + \gamma_j A_j + \delta_{ij} E_i A_j \\ &= \alpha + \\ &\quad + \beta_1 \mathbb{1}_{i=1} + \\ &\quad + \gamma_1 \mathbb{1}_{j=1} + \gamma_2 \mathbb{1}_{j=2} + \gamma_3 \mathbb{1}_{j=3} + \gamma_4 \mathbb{1}_{j=4} + \\ &\quad + \delta_{11} \mathbb{1}_{i=1}\mathbb{1}_{j=1} + \delta_{12} \mathbb{1}_{i=1}\mathbb{1}_{j=2} + \delta_{13} \mathbb{1}_{i=1}\mathbb{1}_{j=3} + \delta_{14} \mathbb{1}_{i=1}\mathbb{1}_{j=4}, \end{aligned}
\end{cases} \tag{1}
$$

where

$$
\mathbb{1}_{x=a} = \begin{cases} 1, & \text{if } x = a \\ 0, & \text{if } x \neq a \end{cases}.
$$

Note that $I_{r_{ij}}$ is a random variable (different individuals with same smoke status and age can suffer or not coronary death within the same follow-up). Hence, the aim of fitting model (1) is to estimate the expectation (i.e. mean) of the coronary deaths rate as:

$$
\mathbb{E}(I_{r_{ij}}) = \mathbb{E}\left(\frac{I_{ij}}{\Delta t_{ij}}\right) = \frac{\mathbb{E}\left(I_{ij}\right)}{\Delta t_{ij}} = \frac{\lambda_{ij}}{\Delta t_{ij}} = \exp\left(L(E_i, A_j)\right).
$$

1. Note that the logarithm function in model (1) implies that we are assuming that the incidence rate varies exponentially with age. Is it consistent with Figure 1? Why?

2. Prove that, under model (1), the expected *IRR* for smokers vs nonsmokers, for a given age group j is:

$$
IRR_j = \begin{cases} \exp\left(\beta_1\right), & \text{if } j = 0 \\ \exp\left(\beta_1 + \delta_{1j}\right), & \text{if } j \neq 0 \end{cases}. \tag{2}
$$

3. According to Figure 1, explain what sign (i.e. positive, negative or zero) do you expect for δ_{1j}.

4. Model (1) can be fitted in R as:

```
> mod <- glm(deaths ~ smoker * age + offset(log(personYears)),
+            family = poisson,
+            data = breslow)
```

that provides the following results:

```
> summary(mod)


##
## Call:
## glm(formula = deaths ~ smoker * age + offset(log(personYears)),
```

```
##     family = poisson, data = breslow)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -9.1479     0.7071 -12.937  < 2e-16 ***
## smokeryes            1.7469     0.7289   2.397  0.01654 *
## age45-54             2.3574     0.7638   3.087  0.00203 **
## age55-64             3.8302     0.7319   5.233 1.67e-07 ***
## age65-74             4.6227     0.7319   6.316 2.69e-10 ***
## age75-84             5.2944     0.7296   7.257 3.96e-13 ***
## smokeryes:age45-54  -0.9866     0.7901  -1.249  0.21174
## smokeryes:age55-64  -1.3628     0.7562  -1.802  0.07151 .
## smokeryes:age65-74  -1.4423     0.7565  -1.906  0.05659 .
## smokeryes:age75-84  -1.8470     0.7572  -2.439  0.01471 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  9.3507e+02  on 9  degrees of freedom
## Residual deviance: -2.6645e-15  on 0  degrees of freedom
## AIC: 75.068
##
## Number of Fisher Scoring iterations: 3
```

5. According the the fitted model, complete the following list **(using `\Sexpr` in your `Rnw` document)**:

   - $\hat{\alpha} =$
   - $\hat{\beta}_1 =$
   - $\hat{\gamma}_1 =$
   - $\hat{\gamma}_2 =$
   - $\hat{\gamma}_3 =$
   - $\hat{\gamma}_4 =$
   - $\hat{\delta}_1 =$
   - $\hat{\delta}_2 =$
   - $\hat{\delta}_3 =$
   - $\hat{\delta}_4 =$

6. Use formula 2 and estimates provided by `summary(mod)` to complete the following list **(using `\Sexpr` in your `Rnw` document)**:

   - $\widehat{IRR}_{35\text{-}44} =$
   - $\widehat{IRR}_{45\text{-}54} =$
   - $\widehat{IRR}_{55\text{-}64} =$
   - $\widehat{IRR}_{65\text{-}74} =$

- $\widehat{IRR}_{\text{75-84}} =$

7. Solve again questions 2. and 3. (only (a), (b) and (c)) in this section, but using now model (1), formula 2 and coefficients estimates provided by `summary(mod)`.

   Show your R code here.

4 Discussion

Based on the results of this analysis of the `breslow` data, discuss, in no more than 10 lines, about if smoking is a risk factor for coronary artery disease. Specifically, take into account concepts such as data context, modeling approach, confusion and interaction.

References

[1] Doll R, Hill A.B. Mortality of British doctors in relation to smoking: Observations on coronary thrombosis. National Cancer Institute Monograph. 1966;19:205-268.

[2] Breslow N.E. Cohort Analysis in Epidemiology. In A Celebration of Statistics A.C. Atkinson and S.E. Fienberg (editors). 1985;109-143. Springer-Verlag.

Delivery:

- **Strictly follow the instructions on delivery of assignments, which are detailed in the syllabus of the course.**
- **Deadline: December 8, 2022 at 22:00h.**