Universitat Autònoma de Barcelona

Facultat de Ciències

Assignment 2

Missing data in linear regression models

Statistics in Health Sciences

Sergi Cantón Simó - 1569251
Cèlia Martínez Frago - 1569504
Goretti Pena Lorente - 1566866
Guillermo Raya García - 1568864

Assignment identifier: A2
Group identifier: G04

8/12/2022

Contents

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 ($j = 0, 1, ..., 4$) for 35-44, 45-54, ..., 75-84, respectively)[1] (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 [1] (one column).

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

```
> Age = c(Data$age[1], Data$age[2], Data$age[3], Data$age[4], Data$age[5])
> Nonsmokers_Py = c(Data$personYears[1], Data$personYears[2],
+                   Data$personYears[3],Data$personYears[4], Data$personYears[5])
> Smokers_Py = c(Data$personYears[6], Data$personYears[7],
+                Data$personYears[8],Data$personYears[9], Data$personYears[10])
> Nonsmokers_Cd = c(Data$deaths[1], Data$deaths[2],
+                   Data$deaths[3], Data$deaths[4],Data$deaths[5])
> Smokers_Cd = c(Data$deaths[6], Data$deaths[7], Data$deaths[8],
+                Data$deaths[9],Data$deaths[10])
> Nonsmokers_Ir = c(Ir_Nonsmokers[1], Ir_Nonsmokers[2], Ir_Nonsmokers[3],
+                   Ir_Nonsmokers[4], Ir_Nonsmokers[5])
> Smokers_Ir = c(Ir_smokers[1], Ir_smokers[2], Ir_smokers[3],
+                Ir_smokers[4], Ir_smokers[5])
> IRR_list = c(IRR[1], IRR[2], IRR[3], IRR[4], IRR[5])
> data = data.frame(Age, Nonsmokers_Py, Smokers_Py, Nonsmokers_Cd, Smokers_Cd,
+                   Nonsmokers_Ir, Smokers_Ir, IRR_list)
>
>
> kable(data, 'latex', booktabs = T, row.names = NA,
+   col.names = c('Age', 'Nonsmokers', 'Smokers',
+                 'Nonsmokers', 'Smokers',
```

[1] See Appendix A.1.

```
+                  'Nonsmokers', 'Smokers', 'Incidence Rate Ratio'),
+   caption = 'Data on coronary death rates including incidence
  rate depending on somke status and incidence rate ratio.') %>%
+   kable_styling(latex_options = "HOLD_position") %>%
+   add_header_above(c(" " = 1,
+   "Person-years" = 2,
+   "Coronary deaths" = 2,
+   "Incidence rate" = 2,
+   " " = 1))
```

Table 2: Data on coronary death rates including incidence rate depending on somke status and incidence rate ratio.

	Person-years		Coronary deaths		Incidence rate		
Age	Nonsmokers	Smokers	Nonsmokers	Smokers	Nonsmokers	Smokers	Incidence Rate Ratio
35-44	18790	52407	2	32	0.11	0.61	5.7
45-54	10673	43248	12	104	1.12	2.40	2.1
55-64	5710	28612	28	206	4.90	7.20	1.5
65-74	2585	12663	28	186	10.83	14.69	1.4
75-84	1462	5317	31	102	21.20	19.18	0.9

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

   (a) The incidence rate among smokers was 2.4 coronary deaths a year per 1000 people.

   (b) The incidence rate among nonsmokers was 1.12 coronary deaths a year per 1000 people.

   (c) The incidence rate ratio was 2.14.

   (d) The *p*-value of the Wald test to decide if the incidence rate among smokers is the same than among nonsmokers was 0.01.

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

   Observing the incidence rate for the age range of 45 to 54 years old, we can see that there were more cases of coronary deaths among the exposed group (roughly 2.4 cases a year per 1000 people, to be specific) than among the unexposed or ”control” group (which suffered approximately 1.12 cases a year per 1000 people). Comparing the two values, we see that the smokers had an incidence rate nearly 2.14 times as big as non-smokers, a number that seems to indicate that smokers in this age group had a bigger chance of suffering from coronary death. In order to check if these results are significant, we have performed a Wald test (the null hypothesis being ”the incidence rate of exposed and non-exposed groups are equivalent”), which estimated an approximate p-value of 0.01. Having received a p-value lower than our threshold of 0.05, we believe that the null hypothesis can be rejected: the test suggests that smoking could be associated with an increased incidence rate of coronary death among our selected demographic.

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

   (a) The incidence rate among smokers was 19.18 coronary deaths a year per 1000 people.

   (b) The incidence rate among nonsmokers was 21.2 coronary deaths a year per 1000 people

   (c) The incidence rate ratio was 0.9.

[2] See Appendix A.2.

[3] See Appendix A.3.

(d) The p-value of the Wald test to decide if the incidence rate among smokers is the same than among nonsmokers was 0.63.

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

Observing the incidence rate for the age range of 75 to 84 years old, we can see that there were less cases of coronary deaths among the exposed group (roughly 19.18 cases a year per 1000 people, to be specific) than among the unexposed or "control" group (which suffered approximately 21.2 cases a year per 1000 people). Comparing the two values, we see that the smokers had an incidence rate about 0.9 times as big as non-smokers, a number that seems to indicate that smokers in this age group had a smaller chance of suffering from coronary death. In order to check if these results are significant, we have performed a Wald test (the null hypothesis being "the incidence rate of exposed and non-exposed groups are equivalent"), which estimated an approximate p-value of 0.63. Having received such a large p-value (way over our threshold of 0.05), we believe that the null hypothesis cannot be rejected with certainty: although our data shows that less individuals in the selected demographic suffered from coronary death, it seems too likely that these results were achieved by chance, and thus we could not conclude that smoking had a clear association with the risk of coronary death in this situation.

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.

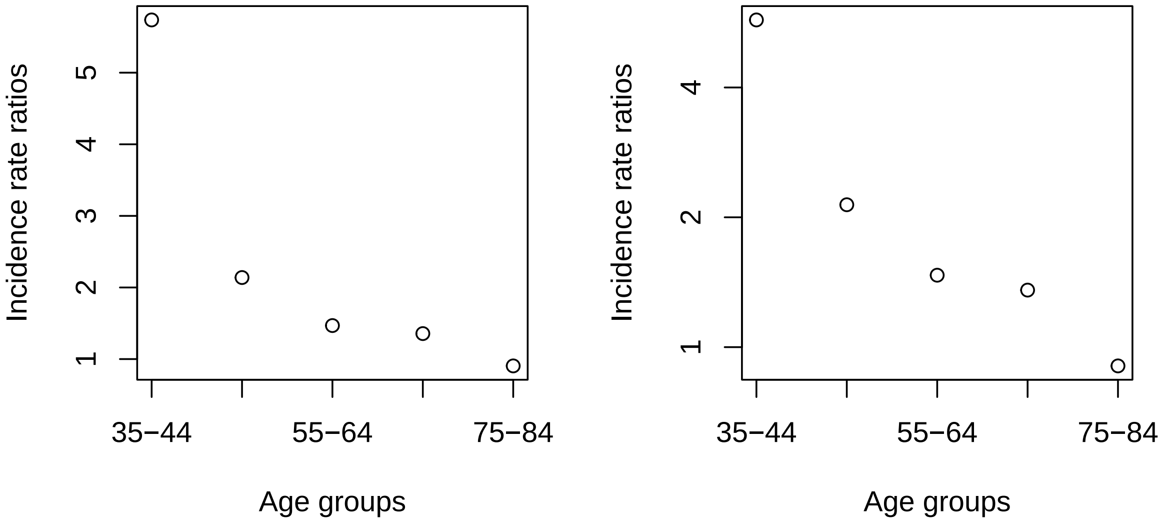

Figure 1: Plots of the incidence rate ratios depending on the age range.The graph on the left is in linear scale, while the graph on the right is on logarithmic scale.

5. According to Figure 1:

   (a) Is the rate ratio greater than 1 for all age groups? What does it mean?
   No, it is not. In both plots, age group 75-84 is less than 1. Such as we know, a rate ratio greater than 1 indicates an increased risk for the group, and a rate ratio less than 1 indicates a decreased risk for the group. So, what it means that 75-84 age group has less risk than the others ages groups.

   (b) Is the rate ratio constant over age groups? What does it mean?
   No, it is not. This means that it is correlated with the age.

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) [4] specified in equations (1) to model the coronary deaths rate for individuals with smoking status i and in age group j, $I_{r_{ij}}$:

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

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:

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

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?

   No, it is not consistent with Figure 1. At first, might seem that on the left plot there is plotted an exponential function. Nevertheless, on the right plot, which is the same function plotted on the log-linear scale, there should be a straight line which clearly it is not.

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

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

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

Proof:

$$\mathbb{E}(IRR_j) = \mathbb{E}\left(\frac{I_{r_{1j}}}{I_{r_{0j}}}\right) = \frac{\mathbb{E}(I_{r_{1j}})}{\mathbb{E}(I_{r_{0j}})} = \begin{cases} (I), & \text{if } j = 0, \\ (II), & \text{if } j \neq 0. \end{cases} \tag{3}$$

Where,

$$(I) = \frac{\exp(\alpha + \beta_1 + 0 + 0)}{\exp(\alpha + 0 + 0 + 0)} = \exp(\alpha + \beta_1 - \alpha) = \exp(\beta_1) \tag{4}$$

$$\begin{aligned} (II) &= \frac{\exp(\alpha + \beta_1 + \sum_{k=1}^{4} \gamma_k \mathbb{1}_{j=k} + \sum_{k=1}^{4} \delta_{1k} \mathbb{1}_{j=k})}{\exp(\alpha + 0 + \sum_{k=1}^{4} \gamma_k \mathbb{1}_{j=k} + 0)} \\ &= \exp(\alpha + \beta_1 + \sum_{k=1}^{4} \gamma_k \mathbb{1}_{j=k} + \sum_{k=1}^{4} \delta_{1k} \mathbb{1}_{j=k} - \alpha - \sum_{k=1}^{4} \gamma_k \mathbb{1}_{j=k}) \\ &= \exp(\beta_1 + \delta_{1j}) \end{aligned} \tag{5}$$

Hence,

$$\mathbb{E}(IRR_j) = \begin{cases} \exp(\beta_1), & \text{if } j = 0, \\ \exp(\beta_1 + \delta_{1j}), & \text{if } j \neq 0. \end{cases} \tag{6}$$

■

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

   In Figure 1 we can see that both graphics are decreasing. That means that, if we increase the age groups (i.e. j value), IRR decreases. If $j = 0$, we do not have any δ_{1j} value, which is equivalent to saying that $\delta_{1j} = 0$. Therefore, δ_{1j} has to be negative when $j > 0$ to make IRR value decrease. In fact, δ_{1j} will have lower values as j becomes larger.

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)           -9.148      0.707  -12.94  < 2e-16 ***
## smokeryes              1.747      0.729    2.40    0.017 *
## age45-54               2.357      0.764    3.09    0.002 **
## age55-64               3.830      0.732    5.23  1.7e-07 ***
## age65-74               4.623      0.732    6.32  2.7e-10 ***
## age75-84               5.294      0.730    7.26  4.0e-13 ***
## smokeryes:age45-54    -0.987      0.790   -1.25    0.212
## smokeryes:age55-64    -1.363      0.756   -1.80    0.072 .
## smokeryes:age65-74    -1.442      0.757   -1.91    0.057 .
## smokeryes:age75-84    -1.847      0.757   -2.44    0.015 *
## ---
## 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: -7.9936e-15  on 0  degrees of freedom
## AIC: 75.07
##
## Number of Fisher Scoring iterations: 3
```

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

   - $\hat{\alpha}$ = -9.15
   - $\hat{\beta}_1 = 1.75$
   - $\hat{\gamma}_1 = 2.36$
   - $\hat{\gamma}_2 = 3.83$
   - $\hat{\gamma}_3 = 4.62$
   - $\hat{\gamma}_4 = 5.29$
   - $\hat{\delta}_1$ = -0.99
   - $\hat{\delta}_2$ = -1.36
   - $\hat{\delta}_3$ = -1.44
   - $\hat{\delta}_4$ = -1.85

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

   - $\widehat{IRR}_{35-44} = 1.75$
   - $\widehat{IRR}_{45-54} = 0.76$
   - $\widehat{IRR}_{55-64} = 0.38$
   - $\widehat{IRR}_{65-74} = 0.3$
   - $\widehat{IRR}_{75-84}$ = -0.1

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

```
> alpha <- mod$coefficients[['(Intercept)']]
> beta1 <- mod$coefficients[['smokeryes']]
> gamma1 <- mod$coefficients[['age45-54']]
> gamma4 <- mod$coefficients[['age75-84']]
> delta1 <- mod$coefficients[['smokeryes:age45-54']]
> delta4 <- mod$coefficients[['smokeryes:age75-84']]
>
>
> Ir_45Nonsmoker_model <- exp(alpha + gamma1)*1000
> Ir_45smoker_model <- exp(alpha + beta1 + gamma1 + delta1)*1000
> Ir_75Nonsmoker_model <- exp(alpha + gamma4)*1000
> Ir_75smoker_model <- exp(alpha + beta1 + gamma4 + delta4)*1000
>
> IRR_45_model <- exp(beta1 + delta1)
> IRR_75_model <- exp(beta1 + delta4)
```

   (a) The estimated values, for the age group of 45 to 54 years old are:

      i. The incidence rate among smokers was 2.4 coronary deaths a year per 1000 people.

      ii. The incidence rate among nonsmokers was 1.12 coronary deaths a year per 1000 people.

iii. The incidence rate ratio was 2.14.

(b) The estimated values, for the age group of 75 to 84 years old are:

i. The incidence rate among smokers was 19.18 coronary deaths a year per 1000 people.

ii. The incidence rate among nonsmokers was 21.2 coronary deaths a year per 1000 people.

iii. The incidence rate ratio was 0.9.

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:

The Wald tests' p-values executed in questions 2.2.e and 2.3.e. suggest a possible association between coronary artery disease and smoking among the younger demographics of our population. Observing Table 2 it is visible that the incidence rate of coronary artery disease increases with age, but the association between said incidence rate and smoking decreases with age too. This makes sense to us because we suspect that older people are more likely to die of coronary death. It should be taken into account, however, that this analysis was performed on data from male doctors, and thus, generalizing the results for the general population might not be accurate. An interesting ampliation of the experiment would be to perform the same tests on a balanced dataset with individuals of different sexes, social backgrounds and economical statuses.

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.

A R code

A.1 Libraries used, data load and, $I_{r_{ij}}$ and IRR_j implementation.

```
> ### libraries used:
> library(knitr)
> library(highlight)  # to highlight R output
> library(xtable)     # to export R output tables to LaTeX
> library(here)
> library(mice)
> library(plotrix)
> library(epitools)
> library(kableExtra)
>
> set.seed(1936)
```

```
> # Ir implementation
> Ir_smokers <- Data$deaths[Data$smoker=='yes']/Data$personYears[Data$smoker=='yes']*1000
> Ir_Nonsmokers <- Data$deaths[Data$smoker=='no']/Data$personYears[Data$smoker=='no']*1000
```

```
> # IRR implementation
> IRR <- Ir_smokers/Ir_Nonsmokers
```

A.2 Implementation of sentences results for age group 45-54.

```
> # The incidence rate among smokers
> Ir_45smoker <- Ir_smokers[2] # coronary deaths a year per 1000 people
>
> # The incidence rate among nonsmokers
> Ir_45Nonsmoker <- Ir_Nonsmokers[2] # coronary deaths a year per 1000 people
>
> # The incidence rate ratio
> IRR_45 <- Ir_45smoker/Ir_45Nonsmoker # rate (no units, since it's a dimensionless magnitude)
>
> # The p-value of the Wald test
> Data45 <- Data[Data$age=="45-54",]
> pValue45 <- rateratio.wald(x=Data45$deaths,y=Data45$personYears,)$p.value[2,"wald"]
```

A.3 Implementation of sentences results for age group 75-84.

```
> # The incidence rate among smokers
> Ir_75smoker <- Ir_smokers[5] # coronary deaths a year per 1000 people
>
> # The incidence rate among nonsmokers
> Ir_75Nonsmoker <- Ir_Nonsmokers[5] # coronary deaths a year per 1000 people
>
> # The incidence rate ratio
> IRR_75 <- Ir_75smoker/Ir_75Nonsmoker # rate (no units, since it is dimensionless)
>
> # The p-value of the Wald test
> Data75 <- Data[Data$age=="75-84",]
> pValue75 <- rateratio.wald(x=Data75$deaths,y=Data75$personYears,)$p.value[2,"wald"]
```

A.4 Implementation of Figure 1.

```
> # Division of space into two plots
> par(mfrow = c(1, 2))
>
> # IRR plot in linear scale
> plot(IRR, xaxt = 'n', xlab = 'Age groups', ylab = 'Incidence rate ratios')
> axis(1, at = seq(1, 5), labels = c('35-44', '45-54', '55-64', '65-74', '75-84'))
>
> # IRR plot in logarithmic scale
> plot(log = 'y', IRR, xaxt = 'n', yaxt = 'n' , xlab = 'Age groups',
+      ylab = 'Incidence rate ratios', ylim = c(min(IRR), max(IRR)))
> axis(1, at = seq(1, 5), labels = c('35-44', '45-54', '55-64', '65-74', '75-84'))
> ylabels <- 2^seq(0, as.integer(log(max(IRR), 2)))
> axis(2, at = ylabels, labels = ylabels)
```