**Clustered Data Practical**

There are two exercises in this practical. Complete exercise 1 in the practical. Exercise 2 can be completed as part of your revision.

|  |
| --- |
| In these exercises we will use the following R functions, for which you should briefly read the documentation:  summary()  skimr::skim()  glm()  jtools::summ()  lmtest::coeftest()  tidyr::pivot\_longer()  tidyr::pivot\_wider()  survival::Surv()  survival::survfit()  plot()  survival::coxph()  You may need to install the following packages:  install.packages("jtools")  install.packages("lmtest")  install.packages("sandwich")  install.packages("skimr")  install.packages("survival")  install.packages("tidyverse")  library(jtools)  library(lmtest)  library(sandwich)  library(skimr)  library(survival)  library(tidyverse) |

Exercise 1: Dataset on birth-weight and maternal smoking

We will analyse a (simulated) retrospective cohort study based on observational data from the Scottish Morbidity record 1992-2006. The Scottish Morbidity Record (SMR02) collects information on clinical and demographic characteristics and outcomes of all patients discharged from Scottish maternity hospitals. We are interested in using the dataset to explore the association between maternal smoking during pregnancy and subsequent low birth weight. The data are stored as a csv dataset called SGA.csv, and the variables are

as follows:

id\_mum maternal number

hospital hospital number 1-48

sga\_5th binary outcome = 1 for small-for-gestational-age birth weight (defined as a birth weight in the smallest 5% for sex and week of gestational age at delivery)

smokcat3 smoking status at the time of first attendance for antenatal care 0=never, 1=smoker, 2=ex smoker

height maternal height in cms

weight maternal weight measured at the time of first attendance for antenatal care

age maternal age defined as the age of the mother at the time of birth

depcat7 carstairs socioeconomic deprivation scores 1(least deprived) to 7 (most deprived)

parcat3 parity was defined as the number of previous livebirths or stillbirths, coded 0=no previous, 1=parity 1-2, 2=parity>=3.

1. Read the data in and familiarise yourself with the structure of the data. Note that each record represents one birth. Are the variables categorical or continuous? How many levels do the categorical variables have? Are there any missing values?

*# Define file path to practical directory (EDIT THIS)*

dir <- "/Users/robertfletcher/Documents/phd/projects"

*# Define practical directory (EDIT THIS)*

prac <- "advanced\_biostats/09\_hierarchical\_data"

*# Read data*

sga <- readr::read\_csv(glue::glue("{dir}/{prac}/data/sga.csv"))

*# Inspect data*

print(sga, n = 10)

skimr::skim(sga)

dplyr::glimpse(sga)

*# Define categorical variables as factors*

sga <- sga |>

dplyr::mutate(

depcat7 = as.factor(depcat7),

hospital = as.factor(hospital),

parcat3 = factor(

parcat3, levels = 1:3,

labels = c("No previous", "Parity 1 to 2", "Parity >=3")

),

sga\_5th = factor(

sga\_5th, levels = 0:1, labels = c("Normal weight", "Low weight")

),

smokcat3 = factor(

smokcat3, levels = 1:3, labels = c("Never", "Smoker", "Ex-smoker")

)

)

2. Explore the structure of the hospital and id\_mum cluster variables. Use length(unique(df$hospital)) and length(unique(df$id\_mum)) to count the number of clusters by hospital and id\_mum.

*# Number of unique hospital clusters*

sga |>

dplyr::summarise(`unique clusters` = length(unique(hospital)))

*# Number of deliveries per hospital*

sga |>

dplyr::count(hospital)

*# Number of births and mothers*

sga |>

dplyr::summarise(births = n(), mothers = length(unique(id\_mum)))

3. Simple logistic analysis of smoking status and SGA.

1. Use logistic regression to explore the association between smoking status and SGA using the following model, and interpret the results:

*# Fit simple logistic regression model*

fit1 <-

glm(sga\_5th ~ smokcat3, data = sga, family = binomial)

fit1 |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

1. Investigate whether the association is confounded by the other maternal characteristics (you may assume linear relationships for height, weight and age and ensure dummy variables are used for for depcat7 and parcat3) .

*# Fit multiple logistic regression model to assess confounding*

fit2 <-

glm(

sga\_5th ~ smokcat3 + height + weight + age + parcat3 + depcat7,

data = sga, family = binomial

)

fit2 |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

4. Obtain robust standard errors accounting for dependence amongst babies born to the same mothers using the command:

lmtest::coeftest(fit1, vcov = sandwich::vcovCL, cluster = sga$id\_mum) Compare these results with those from part 3.

summary(fit1)

5. Obtain robust standard errors dependence amongst babies born in the same hospitals. You could also try adjusting for hospital as shown in lecture. What do you conclude about the extent of clustering by mothers and by hospitals?

*# Robust standard errors for model 1*

lmtest::coeftest(fit1, vcov = sandwich::vcovCL, cluster = sga$hospital)

*# Robust standard errors for model 2*

lmtest::coeftest(fit2, vcov = sandwich::vcovCL, cluster = sga$hospital)

fit3 <-

glm(

sga\_5th ~ smokcat3 + height + weight + age + parcat3 + depcat7 + hospital,

data = sga, family = binomial

) |>

jtools::summ(confint = TRUE, digits = 3, exp = TRUE)

Make sure you understand the difference between the models. Which model do you believe to be more appropriate for these data?

Write a clear statement to summarise the key finding.

Exercise 2: The Diabetic Retinopathy Study (DRS) dataset

We will analyse the Diabetic Retinopathy Study (DRS) dataset. Diabetic retinopathy is the most common and most serious eye complication of diabetes, which may lead to poor vision or even blindness. The DRS was started in 1971 to investigate the efficacy of laser photocoagulation in delaying the onset of severe vision loss. Each patient had one eye randomized to laser treatment and the other eye received no treatment. For each eye, the event of interest was the time from initiation of treatment to the time when visual acuity dropped below 5/200 (defined as "blindness"). Survival times in this dataset are therefore the actual time to blindness in months. Censoring could be caused by death, dropout, or end of the study. The data are stored as DRS.dta and the variables are as follows:

id patient id

age\_dx age at time of diabetes diagnosis

trt1 indicates treatment of left eye coded as 0=untreated, 1=treated

trt2 indicates treatment of right eye coded as 0=untreated, 1=treated

status1 status for left eye coded as 0=censored, 1=blindness

time1 survival time for left eye

status2 status for right eye coded as 0=censored, 1=blindness

time2 survival time for right eye

1. Read the data in and familiarise yourself with the structure of the data. Note that each record represents one individual.

*# Read data*

drs <- readr::read\_csv(glue::glue("{dir}/{prac}/data/drs.csv"))

1. These data are currently in wide format. We need to reshape the data into long format, so that there is one record per eye. Use the following command

*# Convert data to long format (one row per eye)*

drs\_l <- drs |>

tidyr::pivot\_longer(

-c(id, age\_dx), names\_sep = "\\.", names\_to = c("param", "eye")

) |>

tidyr::pivot\_wider(

id\_cols = c(id, age\_dx, eye), names\_from = "param", values\_from = "value"

)

Make sure you understand what these commands has done.

1. Create a survival object

*# Create survival object*

eye\_surv <- survival::Surv(drs\_l$time, drs\_l$blind)

1. Produce a Kaplan meier survival curve for each treatment arm

*# Create Kaplan-Meier object*

km <- survival::survfit(eye\_surv ~ trt, data = drs\_l)

*# Plot Kaplan-Meier curve*

plot(

km, xlab = "months", ylab = "Probability of survival (%)",

main = "Kaplan-Meier survivor function", mark.time = TRUE, yscale = 100

)

(5) Now fit an unadjusted and adjusted cox model to the data assuming all the data are independent and interpret the results

*# Unadjusted*

fit\_un <- survival::coxph(eye\_surv ~ trt, data = drs\_l) |>

summary(exp = TRUE)

*# Adjusted*

fit\_ad <- survival::coxph(eye\_surv ~ trt + eye + age\_dx, data = drs\_l) |>

summary(exp=TRUE)

(6) So far, we have ignored the fact that an individual provides two survival times. The clustering in the coxph function can be incorporated simply using the cluster=variable option in coxph. How do these new results compare with the simple Cox model? What further analyses may be of interest?

*# Unadjusted*

fit\_un2 <-

survival::coxph(eye\_surv ~ trt, cluster = id, data = drs\_l) |>

summary(exp = TRUE)

*# Adjusted*

fit\_ad2 <-

survival::coxph(eye\_surv ~ trt + eye + age\_dx, cluster = id, data = drs\_l) |>

summary(exp = TRUE)