**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)

── Data Summary ────────────────────────

Values

Name sga

Number of rows 6750

Number of columns 11

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Column type frequency:

numeric 11

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Group variables None

── Variable type: numeric ──────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────────

skim\_variable n\_missing complete\_rate mean sd p0 p25 p50 p75 p100 hist

1 height 0 1 162. 6.43 137 158 162 167. 188 ▁▃▇▃▁

2 weight 0 1 3286. 595. 180 2960 3320 3660 9999 ▁▇▁▁▁

3 sex 0 1 1.47 0.499 1 1 1 2 2 ▇▁▁▁▇

4 age 0 1 20.9 3.45 14 18 20 23 32 ▂▇▃▂▁

5 ga 0 1 39.2 2.17 21 38 40 40 42 ▁▁▁▁▇

6 depcat7 0 1 4.68 1.44 1 4 5 6 7 ▂▃▇▆▇

7 parcat3 0 1 1.58 0.607 1 1 2 2 3 ▇▁▇▁▁

8 smokcat3 0 1 1.73 0.632 1 1 2 2 3 ▆▁▇▁▂

9 sga\_5th 0 1 0.0536 0.225 0 0 0 0 1 ▇▁▁▁▁

10 hospital 0 1 19.8 13.2 1 10 17 33 49 ▇▇▃▅▂

11 id\_mum 0 1 1753. 1014. 1 869 1766. 2628 3510 ▇▇▇▇▇

dplyr::glimpse(sga)

> glimpse(sga)

Rows: 6,750

Columns: 11

$ height <dbl> 159, 159, 160, 178, 178, 177, 161, 158, 156, 158…

$ weight <dbl> 2600, 2840, 3140, 3250, 2960, 2760, 3310, 3740, …

$ sex <dbl> 1, 1, 2, 1, 2, 1, 2, 1, 2, 2, 2, 2, 2, 1, 2, 2, …

$ age <dbl> 19, 18, 23, 23, 20, 21, 20, 21, 25, 19, 17, 18, …

$ ga <dbl> 41, 39, 41, 39, 37, 38, 41, 40, 38, 41, 40, 40, …

$ depcat7 <dbl> 4, 4, 4, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 6, 5, 5, …

$ parcat3 <dbl> 2, 2, 3, 2, 1, 2, 1, 2, 3, 2, 1, 1, 2, 1, 1, 2, …

$ smokcat3 <dbl> 2, 2, 2, 3, 2, 2, 3, 2, 2, 2, 2, 2, 3, 2, 1, 3, …

$ sga\_5th <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, …

$ hospital <dbl> 10, 10, 10, 25, 25, 25, 35, 10, 10, 10, 10, 34, …

$ id\_mum <dbl> 1, 1, 1, 2, 2, 2, 3, 4, 4, 4, 4, 5, 6, 6, 7, 7, …

*# 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(hospital)) and length(unique(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)

#There are 48 hospitals of different sizes. There are 6 hospitals with only 1 delivery

#and two hospitals with two deliveries.

*# Number of births and mothers*

sga |>

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

#There are 6750 births from 3510 mothers.

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)

> summ(fit1, confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 6750

Dependent Variable: sga\_5th

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(2) = 83.168, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.036

Pseudo-R² (McFadden) = 0.029

AIC = 2745.235, BIC = 2765.687

Standard errors: MLE

---------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

----------------- ----------- ------- ------- --------- -------

(Intercept) 0.026 0.020 0.033 -28.796 0.000

smokcat32 3.163 2.397 4.174 8.137 0.000

smokcat33 1.590 1.006 2.513 1.984 0.047

---------------------------------------------------------------

> fit1 <- glm(sga\_5th ~ smokcat3, data = df, family = binomial)

> summ(fit1, confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 6750

Dependent Variable: sga\_5th

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(2) = 83.168, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.036

Pseudo-R² (McFadden) = 0.029

AIC = 2745.235, BIC = 2765.687

Standard errors: MLE

---------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

----------------- ----------- ------- ------- --------- -------

(Intercept) 0.026 0.020 0.033 -28.796 0.000

smokcat32 3.163 2.397 4.174 8.137 0.000

smokcat33 1.590 1.006 2.513 1.984 0.047

---------------------------------------------------------------

Smokers have over a three times higher odds of having a small for gestational age baby compared to non-smokers. Ex-smokers have around a 60% higher odds of having a small for gestational age baby compared to non-smokers

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)

> summ(fit2, confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 6750

Dependent Variable: sga\_5th

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(13) = 772.773, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.317

Pseudo-R² (McFadden) = 0.274

AIC = 2077.630, BIC = 2173.072

Standard errors: MLE

--------------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

----------------- ----------- -------- ----------- --------- -------

(Intercept) 2846.640 83.551 96986.854 4.418 0.000

smokcat32 2.526 1.839 3.470 5.719 0.000

smokcat33 1.743 1.043 2.910 2.122 0.034

height 0.962 0.944 0.980 -3.977 0.000

weight 0.998 0.998 0.998 -22.204 0.000

age 1.059 1.011 1.110 2.433 0.015

parcat32 0.551 0.401 0.757 -3.682 0.000

parcat33 0.899 0.521 1.552 -0.381 0.703

depcat72 0.869 0.186 4.070 -0.178 0.859

depcat73 0.786 0.178 3.471 -0.318 0.751

depcat74 1.303 0.304 5.586 0.356 0.722

depcat75 1.128 0.262 4.859 0.161 0.872

depcat76 1.148 0.266 4.967 0.185 0.853

depcat77 1.001 0.228 4.403 0.001 0.999

--------------------------------------------------------------------

The association between smoking status and small for gestational age is confounded slightly by maternal characteristics. The estimated OR is attenuated from 3.16 to 2.53 after adjustment for potential confounders.

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)

> coeftest(fit1, vcov = vcovCL, cluster=df$id\_mum)

z test of coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.64618 0.12873 -28.3234 < 2.2e-16 \*\*\*

smokcat32 1.15149 0.14459 7.9639 1.667e-15 \*\*\*

smokcat33 0.46351 0.24508 1.8913 0.05859 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(fit1)

Call:

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

Deviance Residuals:

Min 1Q Median 3Q Max

-0.3982 -0.3982 -0.2851 -0.2270 2.7100

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.6462 0.1266 -28.796 < 2e-16 \*\*\*

smokcat32 1.1515 0.1415 8.137 4.06e-16 \*\*\*

smokcat33 0.4635 0.2337 1.984 0.0473 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2822.4 on 6749 degrees of freedom

Residual deviance: 2739.2 on 6747 degrees of freedom

AIC: 2745.2

Number of Fisher Scoring iterations: 6

> coeftest(fit2, vcov = vcovCL, cluster=df$id\_mum)

z test of coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 7.95389466 1.89024879 4.2079 2.578e-05 \*\*\*

smokcat32 0.92656700 0.16275229 5.6931 1.247e-08 \*\*\*

smokcat33 0.55537934 0.26342989 2.1083 0.0350083 \*

height -0.03889136 0.01028929 -3.7798 0.0001570 \*\*\*

weight -0.00213407 0.00009415 -22.6667 < 2.2e-16 \*\*\*

age 0.05758392 0.02456038 2.3446 0.0190482 \*

parcat32 -0.59603118 0.16866355 -3.5338 0.0004096 \*\*\*

parcat33 -0.10609198 0.29378459 -0.3611 0.7180085

depcat72 -0.14028446 0.77444185 -0.1811 0.8562556

depcat73 -0.24085176 0.74038958 -0.3253 0.7449509

depcat74 0.26451831 0.73576208 0.3595 0.7192090

depcat75 0.12030476 0.74273884 0.1620 0.8713259

depcat76 0.13841866 0.74116741 0.1868 0.8518507

depcat77 0.00088770 0.75358672 0.0012 0.9990601

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> summary(fit2)

Call:

glm(formula = sga\_5th ~ smokcat3 + height + weight + age + parcat3 +

depcat7, family = binomial, data = df)

Deviance Residuals:

Min 1Q Median 3Q Max

-2.3315 -0.3002 -0.1853 -0.1114 3.4534

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 7.954e+00 1.800e+00 4.418 9.95e-06 \*\*\*

smokcat32 9.266e-01 1.620e-01 5.719 1.07e-08 \*\*\*

smokcat33 5.554e-01 2.617e-01 2.122 0.033804 \*

height -3.889e-02 9.778e-03 -3.977 6.97e-05 \*\*\*

weight -2.134e-03 9.611e-05 -22.204 < 2e-16 \*\*\*

age 5.758e-02 2.367e-02 2.433 0.014987 \*

parcat32 -5.960e-01 1.619e-01 -3.682 0.000231 \*\*\*

parcat33 -1.061e-01 2.784e-01 -0.381 0.703104

depcat72 -1.403e-01 7.877e-01 -0.178 0.858656

depcat73 -2.409e-01 7.578e-01 -0.318 0.750628

depcat74 2.645e-01 7.427e-01 0.356 0.721726

depcat75 1.203e-01 7.452e-01 0.161 0.871741

depcat76 1.384e-01 7.471e-01 0.185 0.853020

depcat77 8.877e-04 7.558e-01 0.001 0.999063

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2822.4 on 6749 degrees of freedom

Residual deviance: 2049.6 on 6736 degrees of freedom

AIC: 2077.6

Number of Fisher Scoring iterations: 7

#Generally the standard errors from the clustered analysis are larger than the simple analysis.

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.

> coeftest(fit1, vcov = vcovCL, cluster=df$hospital)

z test of coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -3.646184 0.096128 -37.9306 < 2e-16 \*\*\*

smokcat32 1.151485 0.127598 9.0243 < 2e-16 \*\*\*

smokcat33 0.463511 0.268575 1.7258 0.08438 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> coeftest(fit2, vcov = vcovCL, cluster=df$hospital)

z test of coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 7.9539e+00 1.6152e+00 4.9244 8.462e-07 \*\*\*

smokcat32 9.2657e-01 1.4973e-01 6.1884 6.078e-10 \*\*\*

smokcat33 5.5538e-01 2.8181e-01 1.9708 0.04875 \*

height -3.8891e-02 9.7874e-03 -3.9736 7.079e-05 \*\*\*

weight -2.1341e-03 8.5818e-05 -24.8675 < 2.2e-16 \*\*\*

age 5.7584e-02 2.1269e-02 2.7074 0.00678 \*\*

parcat32 -5.9603e-01 1.5177e-01 -3.9271 8.598e-05 \*\*\*

parcat33 -1.0609e-01 2.8160e-01 -0.3767 0.70636

depcat72 -1.4028e-01 5.5651e-01 -0.2521 0.80098

depcat73 -2.4085e-01 5.5029e-01 -0.4377 0.66162

depcat74 2.6452e-01 4.9886e-01 0.5302 0.59594

depcat75 1.2030e-01 5.3542e-01 0.2247 0.82222

depcat76 1.3842e-01 5.6134e-01 0.2466 0.80523

depcat77 8.8770e-04 5.4668e-01 0.0016 0.99870

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> fit3 <- glm(sga\_5th ~ smokcat3 + height + weight + age + parcat3 + depcat7 + hospital, data = df, family = binomial)

> summ(fit3, confint = TRUE, digits = 3, exp = TRUE)

MODEL INFO:

Observations: 6750

Dependent Variable: sga\_5th

Type: Generalized linear model

Family: binomial

Link function: logit

MODEL FIT:

χ²(61) = 821.370, p = 0.000

Pseudo-R² (Cragg-Uhler) = 0.335

Pseudo-R² (McFadden) = 0.291

AIC = 2125.033, BIC = 2547.705

Standard errors: MLE

----------------------------------------------------------------------

exp(Est.) 2.5% 97.5% z val. p

----------------- ----------- --------- ------------ --------- -------

(Intercept) 5276.051 138.294 201285.998 4.613 0.000

smokcat32 2.629 1.897 3.642 5.810 0.000

smokcat33 1.767 1.043 2.995 2.115 0.034

height 0.962 0.943 0.981 -3.889 0.000

weight 0.998 0.998 0.998 -22.066 0.000

age 1.042 0.991 1.094 1.617 0.106

parcat32 0.547 0.395 0.757 -3.636 0.000

parcat33 0.885 0.506 1.549 -0.427 0.670

depcat72 0.808 0.170 3.827 -0.269 0.788

depcat73 0.772 0.171 3.482 -0.337 0.736

depcat74 1.374 0.314 6.013 0.422 0.673

depcat75 1.114 0.253 4.903 0.142 0.887

depcat76 1.303 0.292 5.804 0.347 0.729

depcat77 1.212 0.263 5.587 0.247 0.805

hospital2 5.190 1.191 22.611 2.193 0.028

hospital3 0.366 0.065 2.063 -1.139 0.255

hospital4 0.000 0.000 Inf -0.007 0.995

hospital5 1.236 0.482 3.172 0.441 0.659

hospital6 3.381 0.284 40.191 0.965 0.335

hospital7 1.086 0.516 2.283 0.217 0.828

hospital8 0.890 0.482 1.643 -0.372 0.710

hospital9 13.111 1.294 132.865 2.178 0.029

hospital10 0.848 0.493 1.459 -0.595 0.552

hospital11 0.000 0.000 Inf -0.004 0.997

hospital12 0.630 0.332 1.198 -1.408 0.159

hospital13 44.672 3.506 569.215 2.926 0.003

hospital14 0.726 0.261 2.019 -0.613 0.540

hospital15 0.703 0.315 1.568 -0.861 0.389

hospital16 0.594 0.306 1.154 -1.538 0.124

hospital17 0.000 0.000 Inf -0.021 0.983

hospital18 0.977 0.467 2.042 -0.063 0.950

hospital19 0.000 0.000 Inf -0.003 0.997

hospital20 0.497 0.261 0.949 -2.120 0.034

hospital21 0.513 0.208 1.262 -1.454 0.146

hospital22 0.000 0.000 Inf -0.004 0.997

hospital23 0.943 0.437 2.035 -0.150 0.881

hospital24 0.000 0.000 Inf -0.003 0.998

hospital25 1.074 0.612 1.886 0.250 0.802

hospital26 0.000 0.000 Inf -0.016 0.988

hospital27 0.581 0.071 4.733 -0.507 0.612

hospital28 0.000 0.000 Inf -0.005 0.996

hospital29 0.000 0.000 Inf -0.007 0.994

hospital30 0.668 0.083 5.404 -0.378 0.706

hospital31 0.000 0.000 Inf -0.003 0.998

hospital32 10.683 1.281 89.126 2.188 0.029

hospital33 0.694 0.232 2.080 -0.652 0.515

hospital34 0.496 0.190 1.298 -1.428 0.153

hospital35 0.768 0.346 1.704 -0.650 0.516

hospital36 0.000 0.000 Inf -0.005 0.996

hospital37 0.707 0.399 1.254 -1.186 0.236

hospital38 0.430 0.099 1.866 -1.127 0.260

hospital39 1.090 0.139 8.571 0.082 0.935

hospital40 0.000 0.000 Inf -0.014 0.989

hospital41 0.000 0.000 Inf -0.010 0.992

hospital42 0.529 0.196 1.426 -1.258 0.208

hospital43 0.674 0.305 1.488 -0.977 0.329

hospital44 0.000 0.000 Inf -0.018 0.986

hospital45 1.278 0.569 2.871 0.595 0.552

hospital46 2.219 0.449 10.970 0.977 0.328

hospital47 0.000 0.000 Inf -0.003 0.997

hospital48 6.526 1.615 26.365 2.633 0.008

hospital49 0.000 0.000 Inf -0.012 0.990

----------------------------------------------------------------------

#In this example, allowing for clustering within mothers or hospitals altered the standard errors

#(the loss of information has been properly accounted for).

#This had minimal effect on the conclusions – but it was the right thing to do!

#We can only adjust for a single level of clustering using this approach,

#It is best to choose the highest level (ie, the smallest sized clusters if hierarchical ).

#In this case, the smallest clusters are defined by mother\_id.

#Smokers have a 2.5 times higher odds of having a small for gestational age baby compared

# to non-smokers. Ex-smokers have a 1.7 times higher odds of having a small for gestational age baby compared

# to non-smokers.

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"))

> skimr::skim(drs)

── Data Summary ────────────────────────

Values

Name drs

Number of rows 197

Number of columns 8

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Column type frequency:

numeric 8

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Group variables None

── Variable type: numeric ───────────────────────────────────────────────────────────────────────────────

skim\_variable n\_missing complete\_rate mean sd p0 p25 p50 p75 p100 hist

1 id 0 1 873. 496. 5 480 834 1296 1749 ▆▇▇▆▇

2 age\_dx 0 1 20.8 14.8 1 10 16 30 58 ▇▆▃▂▂

3 trt.1 0 1 0.543 0.499 0 0 1 1 1 ▇▁▁▁▇

4 trt.2 0 1 0.457 0.499 0 0 0 1 1 ▇▁▁▁▇

5 blind.1 0 1 0.503 0.501 0 0 1 1 1 ▇▁▁▁▇

6 time.1 0 1 35.5 20.8 0.3 15.8 38.8 54.1 74.9 ▇▅▇▇▅

7 blind.2 0 1 0.274 0.447 0 0 0 1 1 ▇▁▁▁▃

8 time.2 0 1 34.6 21.9 0.83 13.4 38.6 54.1 75.0 ▇▃▆▆▃

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.

> drs\_l

# A tibble: 394 × 6

id age\_dx eye trt blind time

<dbl> <dbl> <chr> <dbl> <dbl> <dbl>

1 5 28 1 1 0 46.2

2 5 28 2 0 0 46.2

3 14 12 1 0 1 42.5

4 14 12 2 1 0 31.3

5 16 9 1 0 0 42.3

6 16 9 2 1 0 42.3

7 25 9 1 1 0 20.6

8 25 9 2 0 0 20.6

9 29 13 1 1 1 0.3

10 29 13 2 0 0 38.8

# … with 384 more rows

# ℹ Use `print(n = ...)` to see more rows

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

)

Chart

Description automatically generated

(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)

> summary(fit\_un, exp=TRUE)

Call:

coxph(formula = eye\_surv ~ trt)

n= 394, number of events= 153

coef exp(coef) se(coef) z Pr(>|z|)

trt -0.2290 0.7954 0.1637 -1.399 0.162

exp(coef) exp(-coef) lower .95 upper .95

trt 0.7954 1.257 0.5771 1.096

Concordance= 0.513 (se = 0.022 )

Likelihood ratio test= 1.98 on 1 df, p=0.2

Wald test = 1.96 on 1 df, p=0.2

Score (logrank) test = 1.97 on 1 df, p=0.2

>

> fit\_ad <- coxph(eye\_surv ~ trt + eye + age\_dx)

> summary(fit\_ad, exp=TRUE)

Call:

coxph(formula = eye\_surv ~ trt + eye + age\_dx)

n= 394, number of events= 153

coef exp(coef) se(coef) z Pr(>|z|)

trt -0.273269 0.760888 0.164432 -1.662 0.0965 .

eye -0.614139 0.541107 0.169871 -3.615 0.0003 \*\*\*

age\_dx 0.004135 1.004143 0.005498 0.752 0.4520

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

trt 0.7609 1.3143 0.5513 1.0502

eye 0.5411 1.8481 0.3879 0.7549

age\_dx 1.0041 0.9959 0.9934 1.0150

Concordance= 0.561 (se = 0.024 )

Likelihood ratio test= 16 on 3 df, p=0.001

Wald test = 15.54 on 3 df, p=0.001

Score (logrank) test = 15.9 on 3 df, p=0.001

(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)

> summary(fit\_un2, exp=TRUE)

Call:

coxph(formula = eye\_surv ~ trt, cluster = id)

n= 394, number of events= 153

coef exp(coef) se(coef) robust se z Pr(>|z|)

trt -0.2290 0.7954 0.1637 0.1362 -1.681 0.0928 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

trt 0.7954 1.257 0.609 1.039

Concordance= 0.513 (se = 0.017 )

Likelihood ratio test= 1.98 on 1 df, p=0.2

Wald test = 2.83 on 1 df, p=0.09

Score (logrank) test = 1.97 on 1 df, p=0.2, Robust = 2.89 p=0.09

(Note: the likelihood ratio and score tests assume independence of

observations within a cluster, the Wald and robust score tests do not).

>

> summary(fit\_ad2, exp=TRUE)

Call:

coxph(formula = eye\_surv ~ trt + eye + age\_dx, cluster = id)

n= 394, number of events= 153

coef exp(coef) se(coef) robust se z Pr(>|z|)

trt -0.273269 0.760888 0.164432 0.130462 -2.095 0.0362 \*

eye -0.614139 0.541107 0.169871 0.140001 -4.387 1.15e-05 \*\*\*

age\_dx 0.004135 1.004143 0.005498 0.006226 0.664 0.5066

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

exp(coef) exp(-coef) lower .95 upper .95

trt 0.7609 1.3143 0.5892 0.9826

eye 0.5411 1.8481 0.4113 0.7120

age\_dx 1.0041 0.9959 0.9920 1.0165

Concordance= 0.561 (se = 0.021 )

Likelihood ratio test= 16 on 3 df, p=0.001

Wald test = 23.92 on 3 df, p=3e-05

Score (logrank) test = 15.9 on 3 df, p=0.001, Robust = 22.81 p=4e-05

(Note: the likelihood ratio and score tests assume independence of

observations within a cluster, the Wald and robust score tests do not).

#Accounting for the dependency in the data due to data coming from the same individuals for each eye is the correct thing to do. In this example, we actually have smaller confidence intervals. This is because we have accounted for the fact that treatment was randomised on each individuals eyes – and so we actually strengthen the evidence.

#In comparison to no treatment, laser eye treatment is associated with a 34% lower risk of severe vision loss.