

```

# FFW2025
# EXERCISES

library(readxl)
library(survey)
library(gee)
library(geepack)
library(lme4)

library(ggplot2)

# LOAD DATA

DATA <- read_excel("responses_2025-09-30_08_50.xlsx", sheet = "Responses")
View(DATA)
summary(DATA)

ggplot(DATA[complete.cases(DATA),], aes(x = CURR_SMOKING, y = DISEASE_COPD, fill = CURR_SMOKING), col = NA) +
  geom_col()
round(prop.table(table(DATA$CURR_SMOKING, DATA$DISEASE_COPD), margin = c(1))*100,1)

# USE COMPLETE DATA
DATA <- DATA[complete.cases(DATA),]

# RECODE CHARACTERS TO FACTORS
DATA[] <- lapply(DATA, function(x) if (is.character(x)) as.factor(x) else x)

# CALCULATE DESIGN WEIGHTS
DATA$W <- 1/DATA$Probs
DATA$W <- DATA$W/sum(DATA$W)*nrow(DATA)

# ANALYSIS 0 : NO ADJUSTMENT

prop.table(table(DATA$DISEASE_COPD))

p <- prop.table(table(DATA$DISEASE_COPD))[2] |> print()
se <- sqrt(p*(1-p)/nrow(DATA)) |> print()

mod0 <- glm(DISEASE_COPD~CURR_SMOKING, family = quasibinomial(link = logit), data = DATA)
summary(mod0)

# ANALYSIS 1 : WEIGHTING, NO ADJUSTMENT OF SE

mod1 <- glm(DISEASE_COPD~CURR_SMOKING, family = quasibinomial(link = logit), weights = W, data = DATA)
summary(mod1)

# ANALYSIS 2: WEIGHTING, ADJUSTED SE (SANDWICH)
# svydesign(ids, strata = NULL, weights=NULL, fpc=NULL, nest = FALSE, data = NULL)

```

```

DESIGN <- svydesign(ids = ~TOWN, strata = NULL, weights=~W, nest = TRUE, fpc = NULL, data = DATA)

svymean(~DISEASE_COPD, design = DESIGN)
confint(svymean(~DISEASE_COPD, design = DESIGN))

svyciprop(~DISEASE_COPD, design = DESIGN)

smod <- svyglm(DISEASE_COPD~CURR_SMOKING, family = quasibinomial(link = logit), design = DESIGN)
summary(smod)

# ANALYSIS 3: GEE

DATA$COPD <- as.numeric(DATA$DISEASE_COPD) - 1
DATA<- DATA[order(DATA$TOWN),]

# UNWEIGHTED

gmod1 <- gee(COPD~CURR_SMOKING, id = TOWN, family = binomial(link = logit), data = DATA, constr = "independence")
summary(gmod1)

gmod2 <- gee(COPD~CURR_SMOKING, id = TOWN, family = binomial(link = logit), data = DATA, constr = "exchangeable")
summary(gmod2)

# WEIGHTED

gmod3 <- geeglm(COPD~CURR_SMOKING, id = TOWN, family = binomial(link = logit), weights = DATA$W , data = DATA, constr =
"independence")
summary(gmod3)

gmod4 <- geeglm(COPD~CURR_SMOKING, id = TOWN, family = binomial(link = logit), weights = DATA$W , data = DATA, constr =
"exchangeable")
summary(gmod4)

# ANALYSIS 3: MULTILEVEL

cmmod <- glmer(COPD ~ CURR_SMOKING + (1|TOWN), family = binomial(link = logit), weights = W, data = DATA)
summary(cmmod)

# POST STRATIFICATION / CALIBRATION

load("population.RData")

P <- droplevels(subset(P, AGE >= 15))
P$AGECAT <- cut(P$AGE, breaks = c(15,25,35,45,55,65,75,120), right = FALSE)
prop.table(table(P$AGECAT))
prop.table(table(P$SEX))

DATA$AGECAT <- cut(DATA$AGE, breaks = c(15,25,35,45,55,65,75,120), right = FALSE)

```

```

prop.table(table(DATA$AGECAT))
prop.table(table(DATA$SEX))

# PREPARE POPULATION DATA

POP_SEX <- data.frame(
  SEX = c("Male", "Female"),
  Freq = c(84813, 82804)
)
POP_AGECA <- data.frame(
  AGECA = c("[15,25)", "[25,35)", "[35,45)", "[45,55)", "[55,65)", "[65,75)", "[75,120)"),
  Freq = c(38744, 40394, 34374, 24193, 16419, 8962, 4531)
)

POP <- list(POP_SEX, POP_AGECA)

# MODMATRIX <- model.matrix(formula, data = P)
# TOTALS <- colSums(MODMATRIX)
# TOTALS <- apply(model.matrix(formula, data = P), 2, sum)

# CALIBRATE

DESIGN <- svydesign(ids = ~TOWN, strata = NULL, weights=~1/Probs, nest = TRUE, fpc = NULL, data = DATA)

# Calibration with raking
raking <- calibrate(DESIGN,
  formula = list(~SEX , ~AGECA),
  population = POP,
  calfun = "raking")
DATA$pw_raking <- weights(raking)

# Linear calibration
linear <- calibrate(DESIGN,
  formula = list(~SEX , ~AGECA),
  population = POP,
  calfun = "linear")
DATA$pw_linear <- weights(linear)

# Calibración logit (
logit <- calibrate(DESIGN,
  formula = list(~SEX , ~AGECA),
  population = POP,
  calfun = "logit", bounds = c(0.3, 3))
DATA$pw_logit <- weights(logit)

# COMPARE

summary(mod0)$coefficients
summary(mod1)$coefficients

```