Untitled

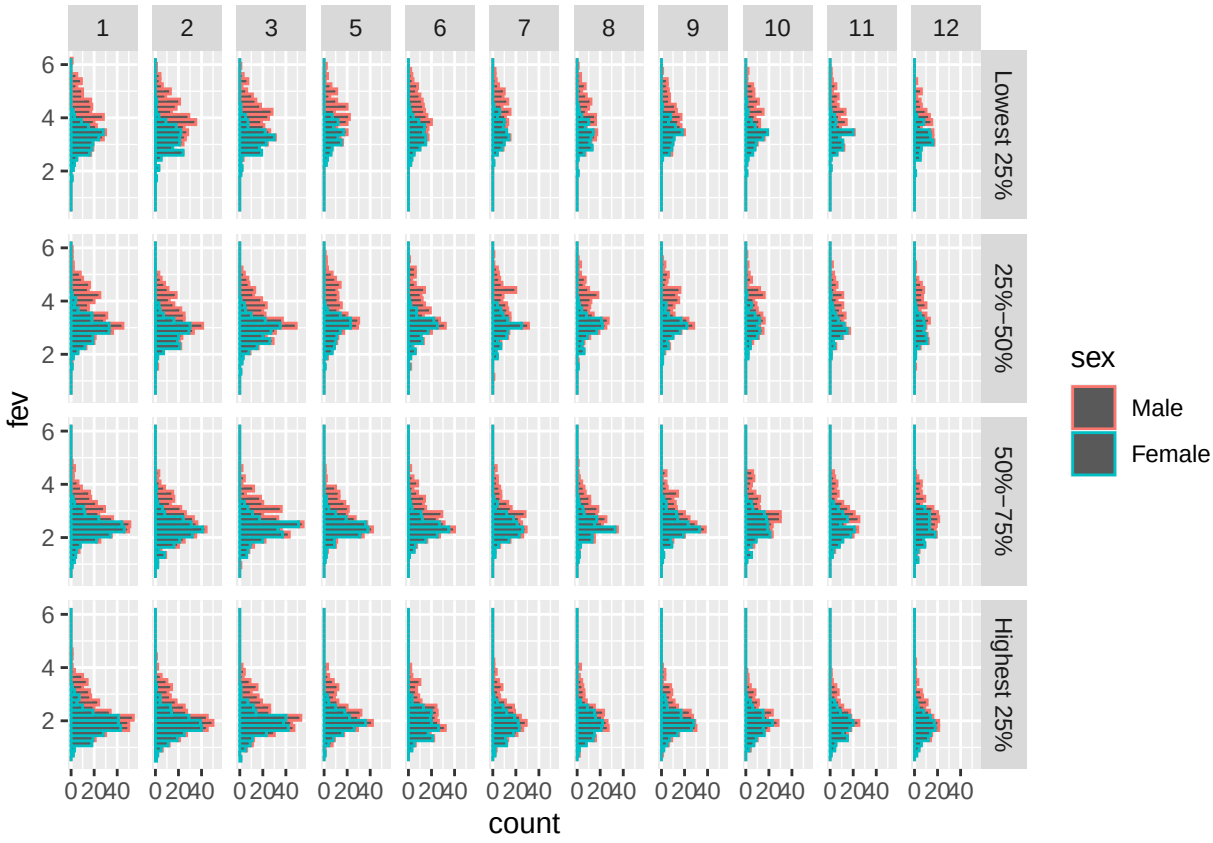
Max Anthenelli

2024-02-21

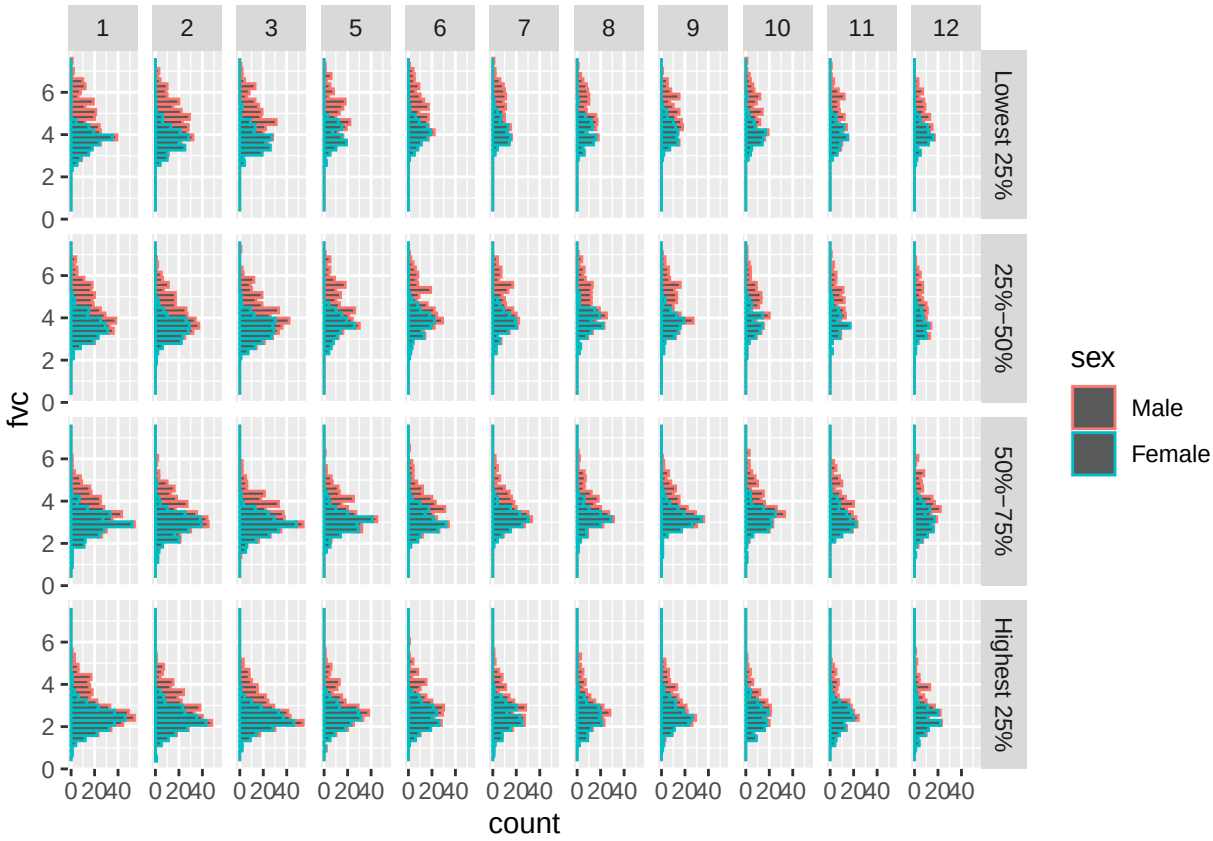
Cleaning procedure

```
knitr::opts_chunk$set(echo = TRUE)
## set working directory ----
working_directory <- file.path("C:", "Users", "Max", "Documents", "GitHub",
                               "STAT741")
setwd(working_directory)

## choose required packages ----
c("tidyverse", "magrittr", "rlang", "glue",
  "haven", "labelled",
  "lme4", "steps", "gt", "gtsummary", "webshot2",
  "scales", "grid", "ggtext", "patchwork"
) |>
purrr::walk(~require(.x, character.only = TRUE))


family_asthma <- read_dta('family_asthma.dta')

long_vars <- c('age', 'smk', 'cncig', 'fvc', 'fev')

pivot_long_repeat_vars <-
  function(var){
  df_name <- glue("family_asthma_long_{var}")
  message(glue("reshaping {df_name}"))
  var_name <- ensym(var)
  df <-
    family_asthma %>%
    pivot_longer(cols = starts_with(var),
                 names_to = 'year',
                 names_prefix = var,
                 values_to = var,
                 values_drop_na = TRUE,
                 names_repair = 'unique') %>%
  select(id, year, all_of(var))
  assign(df_name, df, envir = .GlobalEnv)
  }

walk(long_vars, ~pivot_long_repeat_vars(.x))

family_asthma_long <-
  family_asthma_long_fev %>%
```

```
  full_join(family_asthma_long_cncig, by = c("id", "year")) %>%
  full_join(family_asthma_long_smk, by = c("id", "year")) %>%
  filter(!year =="ever") %>%
  full_join(family_asthma_long_fvc, by = c("id", "year")) %>%
  full_join(family_asthma_long_age, by = c("id", "year")) %>%
  left_join(family_asthma %>% rename(smoke_ever = smkever) %>%
  select(!starts_with(long_vars)), by = c("id")) %>%
  mutate(year =  parse_number(year),
         smk =  ifelse(smk == 1, 1, 0),
         sex = sex - 1,
         sex = as_factor(sex) %>% recode(`0`= "Male", `1` = "Female"),
         cncig = ifelse(smoke_ever == 0, 0 , cncig),
         across(c(id, family), ~to_factor(.x))) %>%
  ungroup() %>%
  group_by(year) %>%
  mutate(cncig_quartile =
          quantile(cncig[cncig>=1], na.rm = TRUE, probs = c(0.25, .5, 0.75)) %>%
          c(-Inf, ., Inf) %>%
          findInterval(cncig, .) %>%
          as_factor() %>%
          recode(`1` = "Lowest 25%", `2` = "25%-50%",
          `3` = "50%-75%", `4` = "Highest 25%", .default = "Didn't Smoke"),
         age_quartile =
           quantile(age, na.rm = TRUE, probs = c(0.25, .5, 0.75)) %>%
          c(-Inf, ., Inf) %>%
          findInterval(age, .) %>%
          as_factor() %>%
          recode(`1` = "Lowest 25%", `2` = "25%-50%",
          `3` = "50%-75%", `4` = "Highest 25%", .default = "Missing Age")) %>%
  ungroup() %>%
  set_variable_labels(
    year = "Time (years)",
    asthma = "Asthmatic",
    mht = "Average Height",
    smoke_ever = "Ever Smoked",
    smk = "Smoke this year",
    age_quartile = "Age Quartile",
    cncig_quartile = "Smoker Frequency Quantile")



family_asthma_long %>%
  filter(!is.na(age_quartile)) %>%
  ggplot(aes(fev)) +
  geom_histogram(aes(color = sex, group = sex)) +
  facet_grid(rows = vars(age_quartile), cols = vars(year)) +
  coord_flip()
```




```
family_asthma_long %>%
  filter(!is.na(age_quartile)) %>%
  ggplot(aes(fvc)) +
  geom_histogram(aes(color = sex, group = sex)) +
  facet_grid(rows = vars(age_quartile), cols = vars(year)) +
  coord_flip()
```



The primary research questions are as follows: 1. Do asthmatics have steeper rates of decline (slope) or lower levels of lung function (FEV1) than non-asthmatics, independent of smoking history?

```
regression_table <- function(object) {
  object %>%
  tbl_regression(
    estimate_fun = partial(style_ratio, digits = 3),
    style_pvalue = function(x) style_pvalue(x, digits =1),
    conf.int = FALSE) %>%
  add_global_p() %>%
  add_significance_stars(
    hide_se = TRUE,
    pattern = "{estimate}{stars}  \n({std.error})"
  ) %>%
  modify_header(estimate ~ "**Beta  \n(SE)**") %>%
  modify_footnote(estimate ~ "SE = Standard Error", abbreviation = TRUE)
}

model_1 <-
  family_asthma_long %>%
  lmer(fev ~ year * asthma + # Main Coefficients of Interest
       mht + sex  + age_quartile + # independent controls
       (1 | family / id) + (age | id) , data = .) %>%
  regression_table()
```

2. Do asthmatic smokers have steeper rates of decline or lower levels of lung function than non-asthmatic smokers? Here you can explore three options 1) using the fixed effect variable ever smoke, 2) using

the time dependent variable for smoking, or 3) using the current number of cigarettes per day. When using this I would suggest making it a categorical variable using quartiles etc.

If you interpret this literally, you should only look at smokers, but I don't think she actually wants that. Should ask tomorrow...

```
model_2 <-
  family_asthma_long %>%
  filter(smoke_ever == 1) %>%
  lmer(fev ~ year * asthma + # Main Coefficients of Interest
       mht + sex  + age_quartile + # independent controls
       (1 | family / id) + (age | id) , data = .) %>%
  regression_table()

model_3 <-
  family_asthma_long %>%
  filter(smoke_ever == 1) %>%
  lmer(fev ~ year * asthma + smk + # Main Coefficients of Interest
       mht + sex  + age_quartile + # independent controls
       (1 | family / id) + (age | id) , data = .) %>%
  regression_table()

model_4 <-
  family_asthma_long %>%
  filter(smoke_ever == 1) %>%
  lmer(fev ~ year * asthma + cncig_quartile + # Main Coefficients of Interest
       mht + sex  + age_quartile + # independent controls
       (1 | family / id) + (age | id) , data = .) %>%
  regression_table()

tbl_merge(
    tbls = map(glue("model_{1:4}"), ~as.symbol(.x) %>% eval_tidy()),
    tab_spanner = glue("**Model {1:4}**")
  )
```

Characteristic	**Beta (SE)**	**Beta (SE)**	**Beta (SE)**	**Beta (SE)**
Time (years)	-0.018*** (0.001)	-0.019*** (0.002)	-0.020*** (0.002)	-0.019*** (0.003)
Asthmatic	-0.204*** (0.052)	-0.087 (0.085)	-0.089 (0.084)	-0.135 (0.085)
Average Height	0.129*** (0.006)	0.109*** (0.010)	0.109*** (0.010)	0.113*** (0.009)
sex	NA*** (NA)	NA*** (NA)	NA*** (NA)	NA*** (NA)
Male	—	—	—	—
Female	-0.342	-0.486	-0.487	-0.465
Age Quartile	NA*** (NA)	NA*** (NA)	NA*** (NA)	NA*** (NA)
Lowest 25%	—	—	—	—
25%-50%	-0.163	-0.168	-0.168	-0.226
50%-75%	-0.354	-0.333	-0.333	-0.422
Highest 25%	-0.465	-0.411	-0.411	-0.536
Time (years) * Asthmatic	0.007 (0.004)	0.004 (0.007)	0.004 (0.007)	0.005 (0.008)
Smoke this year			-0.023 (0.013)	
Smoker Frequency Quantile				NA* (NA)

Characteristic	Beta (SE)	Beta (SE)	Beta (SE)	Beta (SE)
Lowest 25%				—
25%-50%				-0.030
50%-75%				-0.043
Highest 25%				0.011

Hints:

1. When random effects are nested, the order in which they are listed is important. The order in which they are specified (from left to right) is significant – xtmixed assumes that the second factor is nested in the first. In this data subjects (id) are nested within Families (family).
2. Subjects should be tested for a RCM with age but families should not, any adjustment for serial correlation with time done at the subject level would also remove it at the family level.

Notes: 1. Both subject (id) and subject by age, and family (family) should be considered as potential random effects (i.e. fit a RCM) 2. Be sure to select the optimal covariance pattern for your random effects. 3. Plot resulting mean curves whenever you can to demonstrate significant effects. 4. It is not necessary to test higher order polynomials in either the fixed or random effects. 5. Be sure to provide residual plots, normal probability plots for any random effects, and to test for serial correlation in residuals. 6. Lung function measures generally need to be adjusted for anthropometric measures height and sex.