Data Cleaning for Skills for Care Paper

Meng Ye

3/7/2024

## Read me

This is the data cleaning file for generating ipw for the CQC locations selecting into Skills for Care data set. The overall steps is as follows

1. Take the anonymized combined CQC-SfC data, generate a dummy variable based on whether SfC establishment\_id is NA or not, select\_in
2. Filter out the Overall rating rows and calculate the ipw at the location level
3. Run the logit regression model of select\_in and calculate the ipw
4. Derive the sub-data frame of and establishment\_id and ipw
5. Import the cleaned SfC prepared in Step 1
6. Merge the ipw data-frame to the cleaned SfC data

## Setting up

library(tidyverse) # package for data cleaning and plotting  
library(readxl) # package for reading excel file  
library(broom) # extracting model summary as data frame  
library(modelsummary) # deriving model tables  
library(scales) # label percent  
library(lubridate) # working with dates  
library(here) # working with diectory  
library(gt) # grammar of tables

# environment setup to run ordered logit properly  
options(contrasts = rep("contr.treatment", 2))

## Import data

The orignial joint data returned by Skills for care with anonymous providers

#import the raw data file  
cqc\_skills <- read\_excel(here("data","SfC\_CQC.xlsx"))

# data set of rows of CQC with cleaned Skills for Care data

#import the cleaned skills for care data  
cleaned\_cqc\_skills <- read\_rds(here("data","cqc\_sfc\_cleaned.rds"))

# check service group (by Skills for Care) distribution   
cleaned\_cqc\_skills |>  
 group\_by(service\_type\_selected) |>  
 summarize(count = n())

## # A tibble: 5 × 2  
## service\_type\_selected count  
## <chr> <int>  
## 1 Domicilary Care 11148  
## 2 Supported Living 2220  
## 3 care home w/ nursing 8202  
## 4 care home w/o nursing 20400  
## 5 <NA> 1578

## Data Cleaning for CQC\_skills combined data set

# prepare the full CQC data for the logit regression  
cqc <- cqc\_skills %>%   
 mutate(select\_in = ifelse(!is.na(establishmentid), 1, 0)) %>%   
 mutate(inherited = ifelse(inherited == "Y", TRUE, FALSE),  
 rating = recode(rating,   
 "Insufficient evidence to rate" = "NA",  
 "Requires improvement" = "Req improv")) %>%   
 # set the order of the values in the factors   
 mutate(form = fct\_relevel(form, "FPO"),  
   
 # assume the order of the ratings as follows but need to double check with the source   
 rating = ordered(rating, levels = c("Inadequate","Req improv", "Good", "Outstanding"))) %>%  
   
 # adding the rating data coded as numerical  
 mutate(rating\_num = case\_when(rating == "Inadequate" ~ 1,  
 rating == "Req improv" ~ 2,  
 rating == "Good" ~ 3,  
 rating == "Outstanding" ~ 4)) %>%   
   
 mutate(category = case\_when(primary\_cat == "Community based adult social care services" ~ "community",  
 primary\_cat == "Residential social care" ~ "residential",  
 TRUE ~ as.character(NA)),  
  
 # deriving year column and dummy variable for before\_covid  
 year = year(publication\_date),  
 Year = factor(year),  
 during\_covid = ifelse(year >= 2020, TRUE, FALSE),  
 before\_covid = ifelse(year <= 2019, TRUE, FALSE))  
  
head(cqc)

## # A tibble: 6 × 18  
## form cic\_type care\_home primary\_cat region service\_group domain rating  
## <fct> <chr> <chr> <chr> <chr> <chr> <chr> <ord>   
## 1 FPO NA Y Residential socia… East … Overall Safe Good   
## 2 FPO NA Y Residential socia… East … Overall Effec… Good   
## 3 FPO NA Y Residential socia… East … Overall Caring Good   
## 4 FPO NA Y Residential socia… East … Overall Respo… Good   
## 5 FPO NA Y Residential socia… East … Overall Well-… Good   
## 6 FPO NA Y Residential socia… East … Overall Overa… Good   
## # ℹ 10 more variables: publication\_date <dttm>, inherited <lgl>,  
## # establishmentid <dbl>, select\_in <dbl>, rating\_num <dbl>, category <chr>,  
## # year <dbl>, Year <fct>, during\_covid <lgl>, before\_covid <lgl>

## derive overall rating sub data frame

cqc\_overall <- cqc %>%  
 filter(domain == "Overall")

## Run logit model and calculate ipw

model\_select <- glm(select\_in ~ form + category + region +   
 publication\_date + inherited + rating\_num,  
 data = cqc\_overall,  
 family = binomial(link = "logit"))  
  
tidy(model\_select)

## # A tibble: 18 × 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -7.20e- 1 9.95e- 1 -0.724 4.69e- 1  
## 2 formCIC 5.31e- 1 2.27e- 1 2.34 1.92e- 2  
## 3 formGOV 2.59e+ 0 1.92e- 1 13.5 1.04e-41  
## 4 formIND -5.19e- 1 7.81e- 2 -6.65 2.92e-11  
## 5 formNA 1.08e- 1 8.18e- 1 0.132 8.95e- 1  
## 6 formNPO 7.82e- 1 5.24e- 2 14.9 2.12e-50  
## 7 categoryresidential -5.24e- 2 3.92e- 2 -1.34 1.81e- 1  
## 8 regionEast of England 1.56e- 1 7.96e- 2 1.95 5.06e- 2  
## 9 regionLondon -3.70e- 1 8.17e- 2 -4.53 5.99e- 6  
## 10 regionNorth East 1.17e- 1 1.07e- 1 1.09 2.74e- 1  
## 11 regionNorth West 1.19e- 1 7.82e- 2 1.52 1.29e- 1  
## 12 regionSouth East 8.48e- 2 7.21e- 2 1.18 2.40e- 1  
## 13 regionSouth West 2.59e- 1 7.85e- 2 3.30 9.57e- 4  
## 14 regionWest Midlands 5.33e- 2 7.93e- 2 0.672 5.01e- 1  
## 15 regionYorkshire and The Humber 1.00e- 1 8.33e- 2 1.20 2.28e- 1  
## 16 publication\_date -1.51e-10 6.00e-10 -0.252 8.01e- 1  
## 17 inheritedTRUE 4.40e- 1 7.86e- 2 5.60 2.13e- 8  
## 18 rating\_num 2.67e- 1 4.24e- 2 6.31 2.85e-10

## Calculate IPW

ipw\_data <- augment\_columns(model\_select,  
 cqc\_overall,  
 type.predict = "response") %>%   
 rename(propensity = .fitted) %>%   
 mutate(ipw = (select\_in / propensity) + (1 - select\_in) / (1 - propensity)) %>%   
 select(establishmentid, ipw) %>%   
 filter(establishmentid != "NA") %>%   
 group\_by(factor(establishmentid)) %>%   
 slice\_head(n = 1) %>%   
 ungroup()  
  
nrow(ipw\_data)

## [1] 6505

## Merge the ipw data into the cleaned, CQC-Skills for care data

cleaned\_sfc\_ipw <- cleaned\_cqc\_skills %>%   
 left\_join(ipw\_data, by = "establishmentid")

## Save data

write\_rds(cleaned\_sfc\_ipw, file = here("data","cleaned\_sfc\_ipw.rds"))

## Data overview - derive CQC-SFC compare tables

### build a data frame to derive the tables

compare\_data <- cqc %>%   
 left\_join(ipw\_data, by = "establishmentid") %>%   
 mutate(SfC = ifelse(select\_in == 1, TRUE, FALSE),  
 weighted\_rating\_num = rating\_num \* ipw) %>%   
 filter(!is.na(form)& !is.na(rating))

sfc\_count <- compare\_data %>%  
 filter(select\_in == TRUE) %>%   
 group\_by(form, rating) %>%  
 summarize(SfC = n(),  
 SfC\_weighted = sum(ipw, na.rm = TRUE)) %>%   
 ungroup() %>%   
 filter(form !="NA")  
  
sfc\_count

## # A tibble: 19 × 4  
## form rating SfC SfC\_weighted  
## <fct> <ord> <int> <dbl>  
## 1 FPO Inadequate 186 530.   
## 2 FPO Req improv 3013 6939.   
## 3 FPO Good 22779 47013.   
## 4 FPO Outstanding 1163 2153.   
## 5 CIC Req improv 28 52.0   
## 6 CIC Good 250 406.   
## 7 CIC Outstanding 28 41.5   
## 8 GOV Inadequate 2 2.21  
## 9 GOV Req improv 164 176.   
## 10 GOV Good 1886 2019.   
## 11 GOV Outstanding 95 101.   
## 12 IND Inadequate 7 27.1   
## 13 IND Req improv 176 608.   
## 14 IND Good 1397 3984.   
## 15 IND Outstanding 61 148.   
## 16 NPO Inadequate 27 50.1   
## 17 NPO Req improv 521 848.   
## 18 NPO Good 6916 10349.   
## 19 NPO Outstanding 386 541.

cqc\_count <- compare\_data %>%  
 group\_by(form, rating) %>%  
 summarize(CQC = n()) %>%   
 left\_join(sfc\_count, by = c("form", "rating")) %>%   
 ungroup() %>%   
 filter(form!="NA") %>%  
 mutate(ratio = SfC/CQC)

## `summarise()` has grouped output by 'form'. You can override using the  
## `.groups` argument.

cqc\_count

## # A tibble: 19 × 6  
## form rating CQC SfC SfC\_weighted ratio  
## <fct> <ord> <int> <int> <dbl> <dbl>  
## 1 FPO Inadequate 494 186 530. 0.377  
## 2 FPO Req improv 7241 3013 6939. 0.416  
## 3 FPO Good 48269 22779 47013. 0.472  
## 4 FPO Outstanding 2123 1163 2153. 0.548  
## 5 CIC Req improv 44 28 52.0 0.636  
## 6 CIC Good 420 250 406. 0.595  
## 7 CIC Outstanding 40 28 41.5 0.7   
## 8 GOV Inadequate 2 2 2.21 1   
## 9 GOV Req improv 177 164 176. 0.927  
## 10 GOV Good 2045 1886 2019. 0.922  
## 11 GOV Outstanding 100 95 101. 0.95   
## 12 IND Inadequate 46 7 27.1 0.152  
## 13 IND Req improv 655 176 608. 0.269  
## 14 IND Good 4072 1397 3984. 0.343  
## 15 IND Outstanding 100 61 148. 0.61   
## 16 NPO Inadequate 45 27 50.1 0.6   
## 17 NPO Req improv 882 521 848. 0.591  
## 18 NPO Good 10422 6916 10349. 0.664  
## 19 NPO Outstanding 535 386 541. 0.721

sfc\_joint <- compare\_data %>%  
 filter(select\_in == TRUE) %>%   
 group\_by(form) %>%  
 summarize(SfC = n(),  
 SfC\_weighted = sum(ipw, na.rm = TRUE))

cqc\_subtotal <- compare\_data %>%  
 group\_by(form) %>%  
 summarize(CQC = n()) %>%   
 left\_join(sfc\_joint, by = "form") %>%   
 ungroup() %>%   
 filter(form!="NA") %>%  
 mutate(ratio = SfC/CQC) %>%   
 mutate(rating = "Sub-total")  
cqc\_subtotal

## # A tibble: 5 × 6  
## form CQC SfC SfC\_weighted ratio rating   
## <fct> <int> <int> <dbl> <dbl> <chr>   
## 1 FPO 58127 27141 56636. 0.467 Sub-total  
## 2 CIC 504 306 500. 0.607 Sub-total  
## 3 GOV 2324 2147 2299. 0.924 Sub-total  
## 4 IND 4873 1641 4766. 0.337 Sub-total  
## 5 NPO 11884 7850 11788. 0.661 Sub-total

### stack two data frames together

compare\_table <- bind\_rows(cqc\_count, cqc\_subtotal) %>%   
 mutate(rating = factor(rating, levels = c("Outstanding", "Good", "Req improv", "Inadequate", "Sub-total"), ordered = TRUE)) %>%   
 arrange(form, rating) %>%   
 mutate(ratio = scales::percent(ratio, scale = 100, accuracy = 0.1)) %>%   
 mutate(SfC\_weighted= scales::label\_number(big.mark = ",", accuracy = 1)(SfC\_weighted),  
 CQC= scales::label\_number(big.mark = ",", accuracy = 1)(CQC),  
 SfC= scales::label\_number(big.mark = ",",accuracy = 1)(SfC)) %>%   
 gt()  
  
  
   
  
compare\_table

| form | rating | CQC | SfC | SfC\_weighted | ratio |
| --- | --- | --- | --- | --- | --- |
| FPO | Outstanding | 2,123 | 1,163 | 2,153 | 54.8% |
| FPO | Good | 48,269 | 22,779 | 47,013 | 47.2% |
| FPO | Req improv | 7,241 | 3,013 | 6,939 | 41.6% |
| FPO | Inadequate | 494 | 186 | 530 | 37.7% |
| FPO | Sub-total | 58,127 | 27,141 | 56,636 | 46.7% |
| CIC | Outstanding | 40 | 28 | 42 | 70.0% |
| CIC | Good | 420 | 250 | 406 | 59.5% |
| CIC | Req improv | 44 | 28 | 52 | 63.6% |
| CIC | Sub-total | 504 | 306 | 500 | 60.7% |
| GOV | Outstanding | 100 | 95 | 101 | 95.0% |
| GOV | Good | 2,045 | 1,886 | 2,019 | 92.2% |
| GOV | Req improv | 177 | 164 | 176 | 92.7% |
| GOV | Inadequate | 2 | 2 | 2 | 100.0% |
| GOV | Sub-total | 2,324 | 2,147 | 2,299 | 92.4% |
| IND | Outstanding | 100 | 61 | 148 | 61.0% |
| IND | Good | 4,072 | 1,397 | 3,984 | 34.3% |
| IND | Req improv | 655 | 176 | 608 | 26.9% |
| IND | Inadequate | 46 | 7 | 27 | 15.2% |
| IND | Sub-total | 4,873 | 1,641 | 4,766 | 33.7% |
| NPO | Outstanding | 535 | 386 | 541 | 72.1% |
| NPO | Good | 10,422 | 6,916 | 10,349 | 66.4% |
| NPO | Req improv | 882 | 521 | 848 | 59.1% |
| NPO | Inadequate | 45 | 27 | 50 | 60.0% |
| NPO | Sub-total | 11,884 | 7,850 | 11,788 | 66.1% |