IFLS_4_5 IFLS_4_5ing

Chenoa Schatzki-McClain

April 7, 2018

```r
library(lubridate)
```

```
## Warning: package 'lubridate' was built under R version 3.4.3
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:base':
##
##     date
```

```r
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:lubridate':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
rm(list=ls())
```

```r
#load data
IFLS_4_5 <- read.csv("C:/Users/cheno/Desktop/IFLS_all/DATA/R datasets/IFLS_4_5.csv")

#remove unneeded variables
vars <- c("marstat", "sex", "pidlink.x", "pidlink.x.x", "pidlink.x.x.x", "pidlink.y", "pidlink.y.y", "p

for (i in 1:length(vars)){
  IFLS_4_5[, vars[i]] <- NULL
}
```

```r
#remove salary top-coding

IFLS_4_5$tk25a1 <- ifelse(IFLS_4_5$tk25a1 == 999999997, NA, IFLS_4_5$tk25a1)
IFLS_4_5$tk25a1_H <- ifelse(IFLS_4_5$tk25a1_H == 999999997, NA, IFLS_4_5$tk25a1_H)


#remove num job losses top coding >> remove 39
IFLS_4_5$tk46c <- ifelse(IFLS_4_5$tk46c == 99 | IFLS_4_5$tk46c == 98, NA, IFLS_4_5$tk46c)
IFLS_4_5$tk46c_H <- ifelse(IFLS_4_5$tk46c_H == 99  | IFLS_4_5$tk46c_H == 98, NA, IFLS_4_5$tk46c_H)
```

```r
#add employment status variable

IFLS_4_5$employed <- c()

for (i in 1:nrow(IFLS_4_5)){

  IFLS_4_5[i, "employed"] <- ifelse(1 %in% as.numeric(IFLS_4_5[i,c("tk01", "tk01a", "tk03", "tk04")]),

}

IFLS_4_5$employed <- as.factor(IFLS_4_5$employed)

#add ever worked var based on employed and tk05

IFLS_4_5$ever_emp <- ifelse(IFLS_4_5$employed == "1:Employed"| IFLS_4_5$tk06a_H == 1, 1, IFLS_4_5$tk05)

IFLS_4_5$ever_emp <- as.factor(IFLS_4_5$ever_emp)


levels(IFLS_4_5$ever_emp) <- c("1:Yes", "2:No")
```

```r
#add employment status variable

IFLS_4_5$employed_H <- c()

for (i in 1:nrow(IFLS_4_5)){

  IFLS_4_5[i, "employed_H"] <- ifelse(1 %in% as.numeric(IFLS_4_5[i,c("tk01_H", "tk01a_H", "tk03_H", "tk

}

IFLS_4_5$employed_H <- as.factor(IFLS_4_5$employed_H)

#add ever worked var based on employed, tk05, and tk06a

IFLS_4_5$ever_emp_H <- ifelse(IFLS_4_5$employed_H == "1:Employed" | IFLS_4_5$tk06a_H == 1, 1, IFLS_4_5$

IFLS_4_5$ever_emp_H <- as.factor(IFLS_4_5$ever_emp_H)


levels(IFLS_4_5$ever_emp_H) <- c("1:Yes", "2:No")
```

```r
#fix age top-coding

IFLS_4_5$age_H <- ifelse(IFLS_4_5$age_H == 998, NA, IFLS_4_5$age_H)
```

```r
#job loss in last 5 years

IFLS_4_5$job_loss <- ifelse(IFLS_4_5$ever_emp == "3:No" | IFLS_4_5$tk46a == "3:No", 2, IFLS_4_5$tk46b)

IFLS_4_5$job_loss_H <- ifelse(IFLS_4_5$ever_emp_H == "3:No" | IFLS_4_5$tk46a_H == "3:No", 2, IFLS_4_5$t

IFLS_4_5$job_loss <- as.factor(IFLS_4_5$job_loss)
IFLS_4_5$job_loss_H <- as.factor(IFLS_4_5$job_loss_H)
```

```r
levels(IFLS_4_5$job_loss) <- levels(IFLS_4_5$ever_emp)
levels(IFLS_4_5$job_loss_H) <- levels(IFLS_4_5$ever_emp)

#wage info available??

#lots of missing wage info, even for workers in private sector
length(IFLS_4_5[which((is.na(IFLS_4_5$tk25a1_H)) & as.numeric(IFLS_4_5$employed_H) == 1),1])
```

```
## [1] 6751
```

```r
length(IFLS_4_5[which((is.na(IFLS_4_5$tk25a1_H)) & as.numeric(IFLS_4_5$tk24a) == 5),1])
```

```
## [1] 611
```

```r
#job type available >>> job type is available for almost all workers >> could serve as a proxy for wage
length(IFLS_4_5[which((is.na(IFLS_4_5$tk24a)) & as.numeric(IFLS_4_5$employed) == 1),1])
```

```
## [1] 31
```

```r
length(IFLS_4_5[which((is.na(IFLS_4_5$tk24a_H)) & as.numeric(IFLS_4_5$employed_H) == 1),1])
```

```
## [1] 11
```

```r
#compute time between job loss and interview date

#construct interview date and job loss date
IFLS_4_5 <- IFLS_4_5 %>%
          mutate(interview_d = mdy(paste0(ivwmth, "/1/", ivwyr)),
                 jl_d_H = ifelse((is.na(tk46dm_H)), NA, (paste0(tk46dm_H, "/1/", tk46dy_H))))
```

```
## Warning: package 'bindrcpp' was built under R version 3.4.3
```

```r
IFLS_4_5$jl_d_H <- mdy(IFLS_4_5$jl_d_H)

#variable for time between survey and job loss
IFLS_4_5$jl_year <- ceiling(ifelse(is.na(IFLS_4_5$jl_d_H), 0, interval(IFLS_4_5$jl_d_H, IFLS_4_5$interv

#recode for job losses beyond 5 years, likely a recording error, address in data section
IFLS_4_5$jl_year <- ifelse(IFLS_4_5$jl_year %in% c(6,7,8,10, 11, 17), 0, IFLS_4_5$jl_year)

#convert to factor
IFLS_4_5$jl_year <- as.factor(IFLS_4_5$jl_year)

#remove unneeded variables

#remove unneeded variables
vars2 <- c("tk01a", "tk03", "tk04", "tk05", "tk06a", "tk07", "tk25a1x", "tk46a", "tk46b", "dl06x", "ar1

for (i in 1:length(vars)){
  IFLS_4_5[, vars[i]] <- NULL
}

#wave should be factor
IFLS_4_5$year <- as.factor(IFLS_4_5$wave)

#region should be factor
IFLS_4_5$sc_code <- as.factor(IFLS_4_5$sc_code)
```

```r
#remove hours top coding
IFLS_4_5$tk22a <- ifelse(IFLS_4_5$tk22a > 120, NA, IFLS_4_5$tk22a)
IFLS_4_5$tk22a_H <- ifelse(IFLS_4_5$tk22a_H > 120, NA, IFLS_4_5$tk22a_H)

IFLS_4_5$jl_restricted <- ifelse(IFLS_4_5$jl_year %in% 1:2, 1, 0)


IFLS_4_5$employed <- ifelse(as.numeric(IFLS_4_5$employed) == 1, 1, 0)
IFLS_4_5$jl <- ifelse(IFLS_4_5$jl_year == "0", "0", "1")
IFLS_4_5$age_sq <- (IFLS_4_5$age)^2
IFLS_4_5$age_sq_H <- (IFLS_4_5$age_H)^2
IFLS_4_5$jl_year <- as.factor(IFLS_4_5$jl_year)
IFLS_4_5$sc_code <- as.factor(IFLS_4_5$sc_code)
IFLS_4_5$wave <- as.factor(IFLS_4_5$wave)


#wage prof
IFLS_4_5$wage_prof <- ifelse(!(is.na(IFLS_4_5$tk25a1)), IFLS_4_5$tk25a1, ifelse(!(is.na(IFLS_4_5$tk26a1

#remove wage outlier >>> optimize this value later on
IFLS_4_5 <- IFLS_4_5 %>%
            filter(!(!(is.na(wage_prof)) & wage_prof > 10000000))

#remove workers with no hours <<< loook at this later
IFLS_4_5 <- IFLS_4_5 %>%
            filter(!(!(is.na(IFLS_4_5$tk22a)) & IFLS_4_5$tk22a == 0))

#remove workers with no wage <<< look at the later
IFLS_4_5 <- IFLS_4_5 %>%
            filter(!(!(is.na(wage_prof)) & wage_prof == 0))

#limit job loss to firing/displacement
IFLS_4_5$jl_2 <- ifelse(IFLS_4_5$jl == "1" & as.numeric(IFLS_4_5$tk46m_H) %in% c(1, 2, 5), 1, 0)

#limit jl_year based on firings
IFLS_4_5$jl_year_fired <- ifelse(IFLS_4_5$jl_year != "0" & !(as.numeric(IFLS_4_5$tk46m_H) %in% c(1, 2,
IFLS_4_5$jl_year_fired <- as.factor(IFLS_4_5$jl_year_fired)

#eliminate unpaid family workers from employed (employed2)

IFLS_4_5$employed2 <- ifelse(IFLS_4_5$employed == 1 & as.numeric(IFLS_4_5$tk24a) == 6, 0, IFLS_4_5$empl
IFLS_4_5$employed2_H <- ifelse(as.numeric(IFLS_4_5$employed_H) == 1 & as.numeric(IFLS_4_5$tk24a_H) == 6

IFLS_4_5$pidlink <- as.factor(IFLS_4_5$pidlink)

#new job cat
cats <- c("2:self-employed", "2:self-employed", "3:informal business", "4:government", "5:private", "1:u

IFLS_4_5$job_cat_H <- c()

for(i in 1:nrow(IFLS_4_5)){
    if(IFLS_4_5$employed_H[i] == 0){
        IFLS_4_5$job_cat_H[i] <- "1:unemployed"
    }
    else {
```

```r
        IFLS_4_5$job_cat_H[i] <- cats[as.numeric(IFLS_4_5$tk24a_H[i])]
    }
}

IFLS_4_5$job_cat <- c()

for(i in 1:nrow(IFLS_4_5)){
    if(IFLS_4_5$employed[i] == 0){
        IFLS_4_5$job_cat[i] <- "1:unemployed"
    }
    else {
        IFLS_4_5$job_cat[i] <- cats[as.numeric(IFLS_4_5$tk24a[i])]
    }
}

IFLS_4_5$job_cat <- as.factor(IFLS_4_5$job_cat)
IFLS_4_5$job_cat_H <- as.factor(IFLS_4_5$job_cat_H)

write.csv(IFLS_4_5, file = "C:/Users/cheno/Desktop/IFLS_all/DATA/R datasets/IFLS_4_5_4_10.csv")
```