data_stats

Isabella Lin and Sydney Gu

2025-03-10

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

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

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

## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 4.1-8

##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

```
purl("cleaning.Rmd", output = "cleaning2.R")
```

```
##
##
## processing file: cleaning.Rmd

##   |                                                                      |

## output file: cleaning2.R

## [1] "cleaning2.R"
```

```r
source("cleaning2.R")
```

```
## Warning: package 'caret' was built under R version 4.3.3

## Loading required package: ggplot2

## Loading required package: lattice

## Thank you for using fastDummies!

## To acknowledge our work, please cite the package:

## Kaplan, J. & Schlegel, B. (2023). fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from
```

Remove observations with NAs:

```r
cex <- na.omit(cex)
```

Calculate summary statistics:

```r
# CEX
agec <- summary(cex$age)
agecv <- var(cex$age)
famc <- summary(cex$fsize)
famcv <- var(cex$fsize)
incc <- summary(cex$income)
inccv <- var(cex$income)
foodc <- summary(cex$food)
foodcv <- var(cex$food)

# PSID
agep <- summary(psid$age)
agepv <- var(psid$age, na.rm=TRUE)
famp <- summary(psid$fsize)
fampv <- var(psid$fsize)
incp <- summary(psid$income)
incpv <- var(psid$income)
foodp <- summary(psid$food)
foodpv <- var(psid$food, na.rm=TRUE)
```

Format into table:

```
data_stats <- data.frame(
  Variable = c("Age", "Family Size", "Income", "Food Expenditure"),
  Mean = round(c(agec[4], famc[4], incc[4], foodc[4]), 2),
  Median = round(c(agec[3], famc[3], incc[3], foodc[3]), 2),
  Min. = round(c(agec[1], famc[1], incc[1], foodc[1]), 2),
  Max. = round(c(agec[6], famc[6], incc[6], foodc[6]), 2),
  Variance = format(c(agecv, famcv, inccv, foodcv), scientific=TRUE, digits=2),
  Mean = round(c(agep[4], famp[4], incp[4], foodp[4]), 2),
  Median = round(c(agep[3], famp[3], incp[3], foodp[3]), 2),
  Min. = round(c(agep[1], famp[1], incp[1], foodp[1]), 2),
  Max. = round(c(agep[6], famp[6], incp[6], foodp[6]), 2),
  Variance = format(c(agepv, fampv, incpv, foodpv), scientific=TRUE, digits=2),
  check.names = FALSE
)
table <- kable(data_stats, caption = "Comparison of Summary Statistics in CEX and PSID") %>%
  kable_styling("striped") %>%
  add_header_above(c(" " = 1, "CEX" = 5, "PSID" = 5)) %>%
  row_spec(0, bold = TRUE) %>%
  kable_styling("striped", full_width = FALSE) %>%
  column_spec(1, width = "2cm") %>%
  column_spec(2, width = ".8cm") %>%
  column_spec(3, width = ".8cm") %>%
  column_spec(4, width = ".8cm") %>%
  column_spec(5, width = ".8cm")%>%
  column_spec(6, width = ".8cm") %>%
  column_spec(7, width = ".8cm") %>%
  column_spec(8, width = ".8cm") %>%
  column_spec(9, width = ".8cm") %>%
  column_spec(10, width = ".8cm") %>%
  column_spec(11, width = ".8cm")

table
```

Table 1: Comparison of Summary Statistics in CEX and PSID

	CEX					PSID				
Variable	**Mean**	**Median**	**Min.**	**Max.**	**Variance**	**Mean**	**Median**	**Min.**	**Max.**	**Variance**
Age	44.60	43	30	65	8.8e+01	42.81	39	15	99	2.8e+02
Family Size	3.71	4	2	18	2.1e+00	2.97	3	1	19	3.2e+00
Income	40448.92	35547	1	301400	6.2e+08	38333.75	24700	0	2412199	2.2e+09
Food Expenditure	3855.26	3601	288	24165	2.9e+06	3021.20	2600	0	85800	5.3e+06