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
```

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

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

##   |                                                                      |

## output file: cleaning2.R

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

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

```
##
## Attaching package: 'kableExtra'

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

## 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)
psid2 <- na.omit(psid)
dim(psid)
```

```
## [1] 15063    66
```

```r
dim(psid2)
```

```
## [1]  0 66
```

Calculate summary statistics:

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

# PSID
agep <- summary(psid$age)
```

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

Format into table:

```
data_stats <- data.frame(
  Variable = c("Age", "Family Size", "Years of Education", "Income", "Food Expenditure"),
  Mean = round(c(agec[4], famc[4], yoec[4], incc[4], foodc[4]), 2),
  Median = round(c(agec[3], famc[3], yoec[3], incc[3], foodc[3]), 2),
  Min. = round(c(agec[1], famc[1], yoec[1], incc[1], foodc[1]), 2),
  Max. = round(c(agec[5], famc[5], yoec[5], incc[5], foodc[5]), 2),
  Variance = format(c(agecv, famcv, yoecv, inccv, foodcv), scientific=TRUE, digits=2),
  Mean = round(c(agep[4], famp[4], yoep[4], incp[4], foodp[4]), 2),
  Median = round(c(agep[3], famp[3], yoep[3], incp[3], foodp[3]), 2),
  Min. = round(c(agep[1], famp[1], yoep[1], incp[1], foodp[1]), 2),
  Max. = round(c(agep[5], famp[5], yoep[5], incp[5], foodp[5]), 2),
  Variance = format(c(agepv, fampv, yoepv, incpv, foodpv), scientific=TRUE, digits=2),
  check.names = FALSE
)
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)
```

Table 1: Comparison of Summary Statistics in CEX and PSID

	CEX					PSID				
Variable	**Mean**	**Median**	**Min.**	**Max.**	**Variance**	**Mean**	**Median**	**Min.**	**Max.**	**Vari**
Age	44.63	43.0	30	52.0	8.8e+01	44.58	40	16	56	2.8e+
Family Size	3.71	4.0	2	4.0	2.1e+00	2.70	2	1	4	2.2e+
Years of Education	3.84	4.0	1	5.0	2.0e+00	13.35	12	0	14	9.7e+
Income	40161.32	35339.0	-70866	51354.5	6.3e+08	33404.32	25000	-86000	47319	1.6e+
Food Expenditure	3861.07	3602.5	288	4745.0	2.9e+06	3440.05	3120	0	4792	6.9e+