Data Management and Analysis Workshop: Part 2

Data analysis

We will walk through some basic data cleaning, analysis, and visualization tasks using per pupil expenditures as a running example. The repository contains all the data and code we need, so you can just follow along! This notebook introduces descriptive statistics and regression.

Set up

```
# Install required packages.
req <- c("tidyverse", "openxlsx", "broom")
new <- req[!(req %in% installed.packages()[, "Package"])]
if (length(new)) install.packages(new)

# Load required packages.
library(dplyr)
```

```
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
```

```
library(tidyr)
library(broom)
library(openxlsx)
```

```
# Identify inputs and outputs.
PWD <- getwd()
DTA <- file.path(PWD, "out", "data.rda")
OUT <- file.path(PWD, "out", "tables.xlsx")
```

```
# Load analysis file.
# This loads the dataset we created into the same namespace.
load(DTA)
head(dta)
```

```
   leaid             name  stname   tot       exp      ppe pct_asian pct_black
1 100005 Albertville City Alabama  5842  59207000 10134.71 0.4621705  4.193769
2 100006  Marshall County Alabama  5758  68866000 11960.06 0.5036471  1.128864
3 100007      Hoover City Alabama 13640 192421000 14107.11 7.1554252 23.453079
4 100008     Madison City Alabama 11804 184180000 15603.19 9.1070823 19.290071
5 100011       Leeds City Alabama  2097  24080000 11483.07 0.7629948 24.034335
6 100012        Boaz City Alabama  2431  28483000 11716.58 0.6581654  2.591526
   pct_hisp pct_white pct_other maj_group
1 52.601849  39.86648  2.875727      hisp
2 26.432789  70.47586  1.458840     white
3  8.541056  54.95601  5.894428     white
4  7.285666  58.42087  5.896306     white
5 15.069146  58.17835  1.955174     white
6 37.885644  55.49157  3.373097     white
```

Descriptive statistics

```
# Summarize per pupil expenditure and racial composition by majority status.
tab1 <- dta |>
  group_by(maj_group) |>
  summarize(
    n = n(),
    across(ppe, list(mean = mean, min = min, max = max), .names = "{.col}_{.fn}"),
    ppe_p25 = quantile(ppe, probs = 0.25),
    ppe_p50 = quantile(ppe, probs = 0.5),
```

```
    ppe_p75 = quantile(ppe, probs = 0.75),
    across(starts_with("pct_"), ~ mean(.x))
  ) |>
  mutate(pct = n / nrow(dta) * 100) |>
  relocate(pct, .after = n) |>
  relocate(c(ppe_p25, ppe_p50, ppe_p75), .after = ppe_min)
```

RQ1: How does per pupil expenditure vary by school district composition?

```
# Set the reference group among categories.
table(dta$maj_group)
```

```
asian black  hisp  none other white
   71  1444  2317  1971   300 11183
```

```
dta$maj_group <- relevel(as.factor(dta$maj_group), ref = "white")
```

```
# Estimate the partial effects of district composition on per pupil expenditure.
f1 <- ppe ~ maj_group
reg1 <- lm(f1, data = dta)
summary(reg1)
```

```
Call:
lm(formula = f1, data = dta)

Residuals:
    Min      1Q  Median      3Q     Max
 -26841  -12966   -8479   -2822 7209159

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)       26841       1087  24.692  < 2e-16 ***
maj_groupasian   -11512      13685  -0.841  0.40023
maj_groupblack   -10049       3214  -3.126  0.00177 **
maj_grouphisp     -5997       2624  -2.286  0.02228 *
maj_groupnone     -3812       2808  -1.357  0.17468
```

```
maj_groupother     1593       6725   0.237  0.81278
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 115000 on 17280 degrees of freedom
Multiple R-squared:  0.0008398, Adjusted R-squared:  0.0005507
F-statistic: 2.905 on 5 and 17280 DF,  p-value: 0.01263
```

RQ2: Do these differences change with the size of the majority group?

```r
# Get the size of the majority group, i.e., the largest share.
dta <- dta |>
  rowwise() |>
  mutate(
    pct_max = max(c_across(starts_with("pct_"))),
    pts_maj = pct_max - 50
  ) |>
  ungroup()
```

```r
# Estimate the partial effects of group majority size on per pupil expenditure.
f2 <- ppe ~ maj_group * pts_maj
reg2 <- lm(f2, data = dta)
summary(reg2)
```

```
Call:
lm(formula = f2, data = dta)

Residuals:
    Min      1Q  Median      3Q     Max
 -29200  -13460   -7593   -2052 7201185

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           14167.84    2694.46   5.258 1.47e-07 ***
maj_groupasian         1538.06   21284.02   0.072  0.94239
maj_groupblack         3028.95    6695.45   0.452  0.65099
maj_grouphisp          8955.10    5200.45   1.722  0.08509 .
maj_groupnone         11901.30    5072.30   2.346  0.01897 *
maj_groupother        13777.04   12974.79   1.062  0.28833
```

```
pts_maj                   412.94      80.35   5.139 2.78e-07 ***
maj_groupasian:pts_maj  -438.52    1094.23  -0.401  0.68861
maj_groupblack:pts_maj  -427.66     209.71  -2.039  0.04143 *
maj_grouphisp:pts_maj   -508.26     176.33  -2.883  0.00395 **
maj_groupnone:pts_maj     42.57     520.31   0.082  0.93479
maj_groupother:pts_maj  -395.70     390.27  -1.014  0.31064
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 114900 on 17274 degrees of freedom
Multiple R-squared:  0.002432,  Adjusted R-squared:  0.001797
F-statistic: 3.829 on 11 and 17274 DF,  p-value: 1.57e-05
```

Save tables

```
# Write tables to workbook.
# Provide a named list to write dataframes to different worksheets.
ws <- list("Summary" = tab1, "RQ1" = tidy(reg1), "RQ2" = tidy(reg2))
write.xlsx(ws, OUT)
```