

main

nwfried

2024-09-05

Initialise libraries

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr    1.5.1
## v ggplot2    3.5.1      v tibble     3.2.1
## v lubridate  1.9.3      v tidyr      1.3.1
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library("ivreg")
```

Translating do-file “explore_newspaper_msa.do” using <https://www.matthieugomez.com/statar/manipulate-data.html> and google/chatgpt to build dataset of withdrawn/completed highways.

```
#Read in first dataset, withdrawn
folder <- "data"
data1 <- read_csv(file.path(folder, "hwys2msa.csv")) %>%
  filter(withdrawal == 1) %>%
  select(msa = smsacode, length, withdrawal)
```

```
## Rows: 69 Columns: 40
## -- Column specification -----
## Delimiter: ","
## chr  (21): name, layer, path, NHGISNAM, NHGISST, NHGISCTY, ICPSRNAM, STATENA...
## dbl  (17): id, withdrawal, length, length_km, DECADE, ICPSRST, ICPSRCTY, ICP...
## lgl   (1): entityfips
## date  (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#Read in second dataset, completed
data2 <- read_csv(file.path(folder, "pr5112msa.csv")) %>%
  filter(!is.na(smsacode), OPEN90 != 0) %>%
  select(msa = smsacode, length = length_in_km, starts_with("OPEN")) %>%
  mutate(withdrawal = 0)
```

```
## Rows: 10313 Columns: 67
## -- Column specification -----
## Delimiter: ","
```

```

## chr (27): ROUTE_NUM, STATE_FIPS, CNTY_FIPS, FIPS, GEO_ID, STATE, PUMA1, UNIQ...
## dbl (39): ROUTE_2, ROUTE_3, LENGTH, STARTPNT_X, STARTPNT_Y, ENDPNT_X, ENDPNT...
## lgl (1): entityfips
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
withdrawncombined <- bind_rows(data1, data2)

#sort
withdrawncombined <- arrange(withdrawncombined, msa, withdrawal)

withdrawncombined <- withdrawncombined %>%
  group_by(msa) %>%
  mutate(has_withdraw = max(withdrawal)) %>%
  ungroup()

withdrawncombined <- withdrawncombined %>%
  group_by(msa, withdrawal) %>%
  summarize(length = sum(length, na.rm = TRUE), .groups = 'drop')

# Calculate total length for each msa
withdrawncombined <- withdrawncombined %>%
  group_by(msa) %>%
  mutate(total_length = sum(length, na.rm = TRUE)) %>%
  ungroup()

# Generate frac_length and handle precision issues
withdrawncombined <- withdrawncombined %>%
  mutate(frac_length = length / total_length,
         frac_length = ifelse(frac_length > 0.9999, 1, frac_length))

# Sort by msa and descending withdrawal
withdrawncombined <- withdrawncombined %>%
  arrange(msa, desc(withdrawal))

# Tagging the first occurrence of each msa
withdrawncombined <- withdrawncombined %>%
  group_by(msa) %>%
  filter(row_number() == 1) %>%
  ungroup()

# Replace frac_length with 0 where withdrawal is 0
withdrawncombined <- withdrawncombined %>%
  mutate(frac_length = ifelse(withdrawal == 0, 0, frac_length))

# Drop columns withdrawal, tag, and length
msahwy <- withdrawncombined %>%
  select(-withdrawal, -length) %>%
  rename(frac_length_withdrawn = frac_length)

# Save the final dataframe to a CSV file
write_csv(msahwy, file.path(folder, "msahwy.csv"))

```

```
# Display the final dataframe
print(msahwy)
```

```
## # A tibble: 213 x 3
##   msa   total_length frac_length_withdrawn
##   <chr>         <dbl>             <dbl>
## 1 0040          54.2                0
## 2 0080          393.                0
## 3 0160         6047.             0.951
## 4 0200          180.                0
## 5 0240          171.                0
## 6 0320          159.                0
## 7 0360          206.                0
## 8 0400           94.6                0
## 9 0440           64.1                0
## 10 0480          177.                0
## # i 203 more rows
```

The rest of the dofile deals with what we did in the fuzzy_match repo. So we can just read in that data:

```
gentzkowcensus <- read_csv(file.path(folder, "MergedGentzkowCensus.csv"))
```

```
## Rows: 637 Columns: 43
## -- Column specification -----
## Delimiter: ","
## chr (28): NHGISPLACE, STATE, NHGISST, PLACE, GISJOIN, NHGISNAM, NHGISST_2, N...
## dbl (14): citypermid, YEAR, DECADE, ICPSRST, ICPSRCTY, ICPSRSTI, ICPSRCTYI, ...
## lgl (1): entityfips
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
#add circulation data
gentzkow <- read_delim(file.path(folder, "30261-0007-Data.tsv")) %>%
  filter(year == 1960) %>%
  select(citypermid, numdailies, circ, circ_polaff_R, circ_polaff_I, circ_polaff_D, circ_polaff_none, c...
gentzkowcensus <- left_join(gentzkowcensus, gentzkow, by = join_by(citypermid)) %>%
  rename(msa = smsacode)
#add population information
population <- read_csv(file.path("data/nhgis0034_csv/nhgis0034_ds94_1970_place.csv")) %>%
  select(GISJOIN, population = CBC001)
```

```
## Rows: 20950 Columns: 22
## -- Column specification -----
## Delimiter: ","
## chr (6): GISJOIN, STATE, STATEA, PLACE, PLACEA, AREANAME
## dbl (2): YEAR, CBC001
## lgl (14): COUNTYA, CTY_SUBA, TRACTA, ENUMDISTA, CMSA, SMSAA, URB_AREAA, BLCK...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Add population information to this dataset and then merge in with highway data. Generate circulation per capita newspaper info.

```
# merge with hwys2 and msahwy
bygis <- join_by(GISJOIN)
```

```

gentzkowcensushwy <- left_join(gentzkowcensus, population, bygis) %>%
  filter(!is.na(numdailies)) %>%
  filter(population!=0)
# Generate new variables for circulation per capita
gentzkowcensushwy <- gentzkowcensushwy %>%
  mutate(
    circ_per_cap = gentzkowcensushwy$circ / population,
    circ_per_cap_r = gentzkowcensushwy$circ_polaff_R / population,
    circ_per_cap_d = gentzkowcensushwy$circ_polaff_D / population,
    circ_per_cap_i = gentzkowcensushwy$circ_polaff_I / population,
    circ_per_cap_none = gentzkowcensushwy$circ_polaff_none / population
  )
bymsa <- join_by(msa)
gentzkowcensushwy <- inner_join(gentzkowcensushwy, msahwy) %>%
  mutate(lpop = log(population)) %>%
  filter(!is.na(circ_per_cap))

```

```

## Joining with `by = join_by(msa)`
#final dataframe will be named news_hwy for ease of reference
news_hwy <- gentzkowcensushwy

```

Now we can do some OLS Regressions:

```

ols1 <- lm(frac_length_withdrawn ~ circ_per_cap, data = news_hwy)
summary(ols1)

##
## Call:
## lm(formula = frac_length_withdrawn ~ circ_per_cap, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3027 -0.2887 -0.2780  0.6429  0.7491
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.30271    0.02407  12.577  <2e-16 ***
## circ_per_cap -0.03230    0.01664  -1.941   0.0529 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4381 on 442 degrees of freedom
## Multiple R-squared:  0.008453, Adjusted R-squared:  0.00621
## F-statistic: 3.768 on 1 and 442 DF, p-value: 0.05287

ols2 <- lm(frac_length_withdrawn ~ circ_per_cap + lpop, data = news_hwy)
summary(ols2)

```

```

##
## Call:
## lm(formula = frac_length_withdrawn ~ circ_per_cap + lpop, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3209 -0.2892 -0.2757  0.6359  0.7563

```

```
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.18551    0.18169   1.021   0.308
## circ_per_cap -0.03138    0.01671  -1.878   0.061 .
## lpop         0.01057    0.01624   0.651   0.516
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4384 on 441 degrees of freedom
## Multiple R-squared:  0.009405, Adjusted R-squared:  0.004912
## F-statistic: 2.093 on 2 and 441 DF, p-value: 0.1245
```

```
ols3 <- lm(frac_length_withdrawn ~ circ_per_cap_r, data = news_hwy)
summary(ols3)
```

```
##
## Call:
## lm(formula = frac_length_withdrawn ~ circ_per_cap_r, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5914 -0.2805 -0.2639  0.6497  0.7361
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.26389    0.02441  10.809 <2e-16 ***
## circ_per_cap_r  0.04753    0.03955   1.202   0.23
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4392 on 442 degrees of freedom
## Multiple R-squared:  0.003256, Adjusted R-squared:  0.001001
## F-statistic: 1.444 on 1 and 442 DF, p-value: 0.2302
```

```
ols4 <- lm(frac_length_withdrawn ~ circ_per_cap + circ_per_cap_r + circ_per_cap_d + circ_per_cap_i + lpop, data = news_hwy)
summary(ols4)
```

```
##
## Call:
## lm(formula = frac_length_withdrawn ~ circ_per_cap + circ_per_cap_r +
##      circ_per_cap_d + circ_per_cap_i + lpop, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5315 -0.2905 -0.2668  0.6294  0.7793
##
## Coefficients:
##           Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.13581    0.18293   0.742   0.458
## circ_per_cap  -0.61895    0.59935  -1.033   0.302
## circ_per_cap_r  0.65874    0.60172   1.095   0.274
## circ_per_cap_d  0.57861    0.59961   0.965   0.335
## circ_per_cap_i  0.64597    0.61904   1.043   0.297
## lpop           0.01289    0.01624   0.793   0.428
```

```
##
## Residual standard error: 0.4375 on 438 degrees of freedom
## Multiple R-squared:  0.01987,    Adjusted R-squared:  0.008681
## F-statistic: 1.776 on 5 and 438 DF,  p-value: 0.1164

summary(lm(frac_length_withdrawn ~ circ_per_cap + circ_per_cap_r + circ_per_cap_d + circ_per_cap_i + lpop))

##
## Call:
## lm(formula = frac_length_withdrawn ~ circ_per_cap + circ_per_cap_r +
##     circ_per_cap_d + circ_per_cap_i + lpop + numdailies, data = gentzkowcensushwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.5391 -0.2931 -0.2650  0.6296  0.7847
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.09627    0.20674   0.466   0.642
## circ_per_cap   -0.59768    0.60214  -0.993   0.321
## circ_per_cap_r  0.64109    0.60382   1.062   0.289
## circ_per_cap_d  0.56017    0.60185   0.931   0.352
## circ_per_cap_i  0.62657    0.62142   1.008   0.314
## lpop           0.01818    0.02072   0.877   0.381
## numdailies     -0.01382    0.03356  -0.412   0.681
##
## Residual standard error: 0.4379 on 437 degrees of freedom
## Multiple R-squared:  0.02025,    Adjusted R-squared:  0.006798
## F-statistic: 1.505 on 6 and 437 DF,  p-value: 0.1747
```

Create measure of place-level relative circulation to regress on this value:

```
news_hwy <- news_hwy %>% group_by(msa) %>%
  arrange(desc(population), .by_group = TRUE) %>%
  mutate(rel_circ = circ_per_cap / first(circ_per_cap)) %>%
  mutate(rel_circ_r = circ_per_cap_r / first(circ_per_cap_r)) %>%
  mutate(rel_circ_d = circ_per_cap_d / first(circ_per_cap_d)) %>%
  mutate(rel_circ_i = circ_per_cap_i / first(circ_per_cap_i)) %>%
  ungroup()
```

Now we can run a few regressions using this relative circulation information:

```
summary(lm(frac_length_withdrawn ~ rel_circ, data = news_hwy))

##
## Call:
## lm(formula = frac_length_withdrawn ~ rel_circ, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3075 -0.2818 -0.2811  0.6428  0.7412
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.30747    0.02336  13.160 < 2e-16 ***
## rel_circ     -0.02634    0.01005  -2.621  0.00908 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4366 on 442 degrees of freedom
## Multiple R-squared:  0.0153, Adjusted R-squared:  0.01307
## F-statistic: 6.868 on 1 and 442 DF,  p-value: 0.009077
summary(lm(frac_length_withdrawn ~ rel_circ + lpop, data = news_hwy))

##
## Call:
## lm(formula = frac_length_withdrawn ~ rel_circ + lpop, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3123 -0.2882 -0.2804  0.6350  0.7597
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.210622   0.181531   1.160   0.2466
## rel_circ     -0.025755   0.010115  -2.546   0.0112 *
## lpop          0.008726   0.016220   0.538   0.5909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4369 on 441 degrees of freedom
## Multiple R-squared:  0.01595,    Adjusted R-squared:  0.01148
## F-statistic: 3.573 on 2 and 441 DF,  p-value: 0.02888
```

Read in adpricing data for IV regression and m:1 merge with news_hwy dataframe.

```
adpricing <- read_csv(file.path("data/ads.csv"))

## Rows: 1247 Columns: 2
## -- Column specification -----
## Delimiter: ","
## dbl (2): citypermid, adprice
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
news_hwy <- left_join(news_hwy, adpricing, join_by(citypermid))
```

IV regress using adprice variable

```
summary(ivreg(frac_length_withdrawn ~ rel_circ | adprice, data = news_hwy))

##
## Call:
## ivreg(formula = frac_length_withdrawn ~ rel_circ | adprice, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -5.8227 -0.2950 -0.2103  0.6716  0.8899
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.08945    0.18770   0.477   0.634
```

```
## rel_circ      0.20551    0.17276    1.190    0.235
##
## Diagnostic tests:
##              df1 df2 statistic p-value
## Weak instruments    1 362      3.226 0.0733 .
## Wu-Hausman          1 361      4.203 0.0411 *
## Sargan              0 NA         NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6799 on 362 degrees of freedom
## Multiple R-Squared:  -1.254, Adjusted R-squared:  -1.26
## Wald test: 1.415 on 1 and 362 DF, p-value: 0.235
summary(ivreg(frac_length_withdrawn ~ rel_circ + lpop | adprice + lpop, data = news_hwy))
```

```
##
## Call:
## ivreg(formula = frac_length_withdrawn ~ rel_circ + lpop | adprice +
##       lpop, data = news_hwy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.6267 -0.3208 -0.2585  0.6194  0.7991
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.10458     0.24938  -0.419   0.675
## rel_circ      0.05153     0.04537   1.136   0.257
## lpop          0.03268     0.02055   1.590   0.113
##
## Diagnostic tests:
##              df1 df2 statistic  p-value
## Weak instruments    1 361    24.854 9.62e-07 ***
## Wu-Hausman          1 360     3.506  0.062 .
## Sargan              0 NA         NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4795 on 361 degrees of freedom
## Multiple R-Squared:  -0.1177, Adjusted R-squared:  -0.1239
## Wald test: 1.423 on 2 and 361 DF, p-value: 0.2424
```

Create dataframe for MSA-level data & regressions:

```
news_hwy_msa <- news_hwy %>% group_by(msa) %>%
  arrange(desc(population), .by_group = TRUE) %>%
  summarise(total_circ_per_cap = sum(circ_per_cap, na.rm = TRUE), city_circ_per_cap = first(circ_per_cap),
  mutate(msa_rel_circ = city_circ_per_cap / total_circ_per_cap)
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## i Please use `reframe()` instead.
## i When switching from `summarise()` to `reframe()`, remember that `reframe()`
## always returns an ungrouped data frame and adjust accordingly.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
```



```
## generated.
## `summarise()` has grouped output by 'msa'. You can override using the `.groups`
## argument.
summary(lm(frac_length_withdrawn ~ msa_rel_circ, data = news_hwy_msa))

##
## Call:
## lm(formula = frac_length_withdrawn ~ msa_rel_circ, data = news_hwy_msa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.4789 -0.3335 -0.1351  0.4815  0.9247
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.49110    0.03658  13.426 < 2e-16 ***
## msa_rel_circ -0.41584    0.06030  -6.897 1.85e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.418 on 442 degrees of freedom
## Multiple R-squared:  0.09716,    Adjusted R-squared:  0.09511
## F-statistic: 47.56 on 1 and 442 DF,  p-value: 1.849e-11
```