Analysis of MSA income outcomes

Ari Anisfeld

February 22, 2019

Set-up

```
raw_borja_2000 <- read_dta("BORJAS2000FINAL.dta")


borja_2000 <- raw_borja_2000 %>%
  select(lincShared, lr, skill, statefip, pumares2mig)

ipums <-read_excel("ipums_usa_puma_migpuma_2000.xlsx")

puma_names <-
  read_table("2000PUMAsASCII.txt",
             col_names = c("blank", "col_letters", "col_strings" ),
             skip = 24, na = c("", "NA")) %>%
    slice(1:15) %>%
    select(-blank) %>%
    filter(!is.na(col_letters))

raw_puma_data <-
  read_table("2000PUMAsASCII.txt",
             col_names = puma_names$col_strings,
             skip = 42, na = c("", "NA"),
             guess_max = 100000)


puma_data <-
  raw_puma_data %>%
  filter(is.na(`Census Tract Code`), !is.na(`FIPS Place Code`)) %>%
  transmute(name = `Area Name`,
            msa = `Metropolitan Statistical Area/Consolidated`,
            population = `Census 2000 100% Population Count`,
            PUMA=`PUMA Code`,
            statefip =as.numeric(`FIPS State Code`))
by_place <-
ipums %>%
  mutate(MIGPUMA =as.numeric(`Migration PUMA, first 3 digits (MIGPUMA)`),
         statefip = as.numeric(`State FIPS Code (STATEFIP)`)) %>%
  left_join(puma_data, by=c("PUMA", "statefip")) %>%
  filter(!str_detect(name, "PUMA [0-9]{5}")) %>%
  group_by(statefip, MIGPUMA, name) %>%
  summarize(population_by_town = sum(population)) %>%
  arrange(desc(population_by_town))


sanity_check <- by_place %>% ungroup() %>% summarize(`us population` = sum(population_by_town))
```

```
us_pop <- sanity_check %>% pull(`us population`)
sanity_check %>% knitr::kable()
```

us population
285230516

Merge characteristics

```
* what fraction of MIGPUMAs contain multiple place names: 97 percent
* what fraction of MIGPUMAs contain multiple MSAs: 14 percent
* when you have a MIGPUMA that contains multiple names, what fraction of the population lives in the pla
    The named places account for 33 percent of the US population.
    The table below shows how that proportion changes as the number of places increase.
```

```
top_line_place <- by_place %>%
    mutate(name = str_replace_all(name, "(Remainder of | \\(part\\)| \\(balance\\))", "")) %>%
    group_by(statefip, MIGPUMA) %>%
    summarize(name = first(name),
              place_count = n(),
              top_place_population = first(population_by_town),
              proportion_in_named_place = top_place_population/sum(population_by_town)) %>%
  ungroup()


top_line_place %>%
  count(place_count, proportion_in_named_place) %>%
  mutate(bins =
           case_when(
             place_count == 1 ~ "a) 1",
             place_count <= 10 ~ "b) 2 - 10",
             place_count <= 20 ~ "c) 11 - 20",
             place_count <= 50 ~ "d) 21 - 50",
             place_count <= 100 ~ "e) 51 - 100",
             TRUE ~ "f) > 100"),
         total = sum(n)
  ) %>%
  group_by(bins) %>%
  summarize(n_migpuma = sum(n),
            proportion_in_bin = n_migpuma / first(total),
            mean_proportion_in_named_place = mean(proportion_in_named_place)) %>%
  knitr::kable(digits=3)
```

bins	n_migpuma	proportion_in_bin	mean_proportion_in_named_place
a) 1	33	0.031	1.000
b) 2 - 10	129	0.123	0.538
c) 11 - 20	228	0.217	0.384
d) 21 - 50	383	0.365	0.263
e) 51 - 100	178	0.170	0.185
f) > 100	99	0.094	0.175

```
top_line_place %>% summarise(percent_coverage_us = sum(top_place_population)/us_pop)
```

```
## # A tibble: 1 x 1
##   percent_coverage_us
##                 <dbl>
## 1               0.366
```

```
borja_2000 %>%
  left_join(top_line_place, by=c("statefip"= "statefip", "pumares2mig" = "MIGPUMA")) %>%
  select(name, skill, lincShared, lr, statefip, pumares2mig) %>%
  write_excel_csv("inc_by_skill_by_place.csv")
```

```
## Warning: Column `statefip` has different attributes on LHS and RHS of join
```

```
## Warning: Column `pumares2mig`/`MIGPUMA` has different attributes on LHS and
## RHS of join
```

Top MSA in terms of low-skilled workers earnings

```
by_msa <-
ipums %>%
  mutate(MIGPUMA =as.numeric(`Migration PUMA, first 3 digits (MIGPUMA)`),
         statefip = as.numeric(`State FIPS Code (STATEFIP)`)) %>%
  left_join(puma_data, by=c("PUMA", "statefip")) %>%
  filter(!str_detect(name, "PUMA [0-9]{5}")) %>%
  group_by(statefip, MIGPUMA, msa) %>%
  arrange(desc(population)) %>%
  summarize(name = first(name),
            population_by_msa = sum(population)) %>%
  arrange(desc(population_by_msa))

top_line_msa <- by_msa %>%
    group_by(statefip, MIGPUMA) %>%
    summarize(msa = first(msa),
              name = first(name),
              place_count = n(),
              top_place_population = first(population_by_msa),
              proportion_in_named_place = top_place_population/sum(population_by_msa)) %>%
  ungroup()


msa_data <-
raw_borja_2000 %>%
  left_join(top_line_msa, by=c("statefip"= "statefip", "pumares2mig" = "MIGPUMA")) %>%
  select(-c(proportion_in_named_place )) %>%
  filter(msa!=9999) %>%
  arrange(desc(top_place_population)) %>%
  group_by(msa, skill) %>%
  summarize(name = first(name),
            pop = sum(top_place_population),
            lincShared = weighted.mean(lincShared, w = basePopTot),
            lr= weighted.mean(lr, w = basePopTot)) %>%
  select(name, skill, everything()) %>%
```

```
  mutate(gap =  lr / lincShared - 1) %>%

  filter(skill == 0) %>%
  arrange(desc(gap))
```

```
## Warning: Column `statefip` has different attributes on LHS and RHS of join
```

```
## Warning: Column `pumares2mig`/`MIGPUMA` has different attributes on LHS and
## RHS of join
```

```
object <-
msa_data %>%
  ggplot(aes(x=lincShared, y=lr)) +
  geom_point() +
  geom_smooth(se = F) +
  theme_minimal() +
  labs(title = "Log wages of unskilled workers compared to weighted MSA average log wage")



model <- loess(lr~lincShared, data=msa_data)

msa_data %>%
  mutate(lr_hat = predict(model, lincShared),
         resid = lr - lr_hat) %>%
 arrange(desc(resid)) %>% select(-skill, -msa, -gap) %>%
  head() %>% knitr::kable(digits = 3)
```

```
## Adding missing grouping variables: `msa`
```

msa	name	pop	lincShared	lr	lr_hat	
3850	Kokomo city (part)	101541	11.269	11.094	10.905	
0460	Oshkosh city	358365	11.256	11.024	10.896	
4240	Lewiston city	90830	11.128	10.911	10.792	
4800	Mansfield city	128852	11.074	10.870	10.755	
4320	Lima city	155084	11.137	10.908	10.798	
3620	Janesville city	152307	11.249	10.999	10.891	
To iden	tify Municiple Statis	tical Are	as (MSA) with	relative	ly good o	utcomes for low-skilled workers, I first

This analysis could be extended by including the consumption data from borja_2000.

```
msa_data %>%
  mutate(lr_hat = predict(model, lincShared),
         resid = lr - lr_hat) %>%
  filter(pop > 1000000) %>% arrange(desc(resid)) %>% select(-skill, msa, gap) %>%
  head() %>% knitr::kable(digits = 3)
```

name	msa	pop	lincShared	lr	gap	lr_hat	resid
Detroit city (part)	2162	5456428	11.380	11.052	-0.029	10.964	0.088
Grand Rapids city	3000	1088514	11.268	10.986	-0.025	10.904	0.081
Minneapolis city (part)	5120	2868847	11.442	11.054	-0.034	10.978	0.076
Henderson city	4120	1530797	11.180	10.894	-0.026	10.833	0.062
Independence city (part)	3760	1679020	11.308	10.976	-0.029	10.930	0.046

name	msa	pop	lincShared	lr	gap	lr_hat	resid
Remainder of Fairfax County (part)	8872	7492944	11.471	11.021	-0.039	10.980	0.041