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

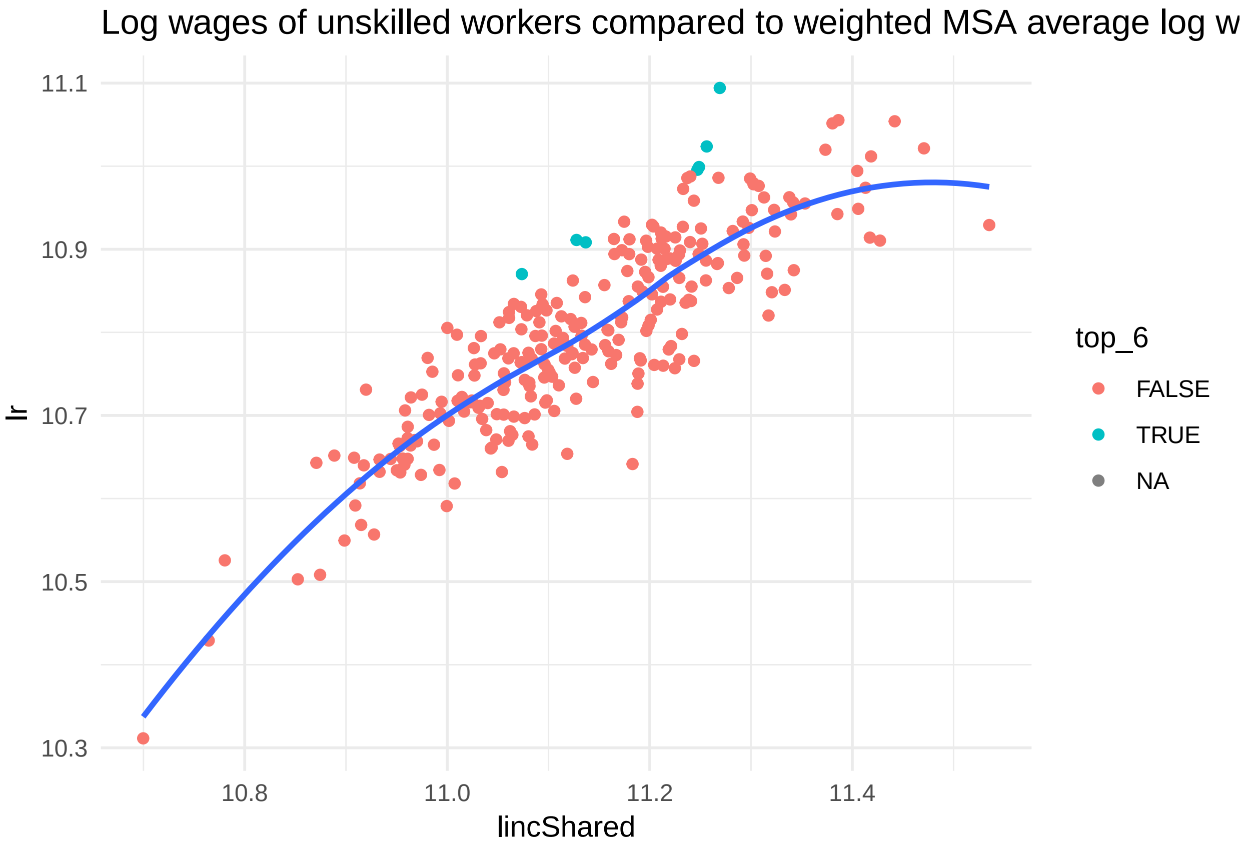
```
top_line_place %>% summarise(percent_coverage_us = sum(top_place_population)/us_pop)
```

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

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))


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

```r
msa_data <-
msa_data %>%
  ungroup() %>%
  mutate(name = str_replace_all(name, "(Remainder of | \\(part\\)| \\(balance\\))", ""),
         lr_hat = predict(model, lincShared),
         resid = lr - lr_hat,
         rank = dense_rank(desc(resid)),
         top_6 = rank <= 7)

msa_data %>%
  ggplot(aes(x=lincShared, y=lr)) +
  geom_point(aes(color=top_6)) +
  geom_smooth(se = F) +
  theme_minimal() +
  labs(title = "Log wages of unskilled workers compared to weighted MSA average log wage")
```



```r
msa_data %>%
 arrange(desc(resid)) %>% select(-skill, -msa, -gap, -top_6) %>%
  head() %>% knitr::kable(digits = 3)
```

name	pop	lincShared	lr	lr_hat	resid	rank
Kokomo city	101541	11.269	11.094	10.905	0.189	1
Oshkosh city	358365	11.256	11.024	10.896	0.128	2
Lewiston city	90830	11.128	10.911	10.792	0.120	3
Mansfield city	128852	11.074	10.870	10.755	0.115	4
Lima city	155084	11.137	10.908	10.798	0.110	5

name	pop	lincShared	lr	lr_hat	resid	rank
Janesville city	152307	11.249	10.999	10.891	0.108	6

Analysis

To identify Municiple Statistical Areas (MSA) with relatively good outcomes for low-skilled workers, I first estimate the expected log-wage for low-skill workers given the average log-wages in the MSA. I then compare the actual outcomes to the expected outcomes. The key metric, "resid" in the table above, is a measure of how much better the MSA performs compared to the estimate. The table shows mostly small towns in the midwest. "resid" is measured in log points and so can be interpreted as a percentage bonus for low-skilled workers in the MSA. For example, in Kokomo there is an 19 percent X relative to an average MSA with the same shared log-wages. In the table below, I restrict the sample to MSAs with over 1 million residents. Of these larger population centers, Detroit is the top by our metric (17th overall), with low-skilled workers earning nearly 9 percent (.09 log-points) more than expected.

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, -top_6) %>%
  head() %>% knitr::kable(digits = 3)
```

name	pop	lincShared	lr	lr_hat	resid	rank
Detroit city	5456428	11.380	11.052	10.964	0.088	17
Grand Rapids city	1088514	11.268	10.986	10.904	0.081	20
Minneapolis city	2868847	11.442	11.054	10.978	0.076	26
Henderson city	1530797	11.180	10.894	10.833	0.062	41
Independence city	1679020	11.308	10.976	10.930	0.046	58
Fairfax County	7492944	11.471	11.021	10.980	0.041	68