EDS 241: Assignment 1

Mia Forsline

1/21/2022

1 Introduction

This assignment uses data from CalEnviroScreen 4.0, a mapping and data tool produced by the California Office of Environmental Health Hazards Assessment (OEHHA). The data are compiled and constructed from a variety of sources and cover all 8,035 census tracts in California.

Specifically, I used the following variables:

- census tract ID,
- total population per census tract,
- California county name (the county where the census tract is located),
- Low Birth Weight (% of census tract births with weight $<$ 2500g),
- PM25 (ambient concentrations of PM2.5 in the census tract, in µg/m^3),
- and Poverty (% of population in the census tract living below twice the federal poverty line).

2 Read in and clean data

Select variables of interest

```
data <- read_csv(here("data", "CES4.csv"))

data_clean <- data %>%
      clean_names()


data_clean <- data_clean %>%
      select(census_tract,
             california_county,
             total_population,
             low_birth_weight,
             pm2_5,
             poverty)
```

3 (a) What is the average concentration of PM2.5 across all census tracts in California?

```
mean_pm2.5 <- mean(data_clean$pm2_5) %>%
  round(digits = 2)
```

The mean concentration of PM2.5 across all census tracts in California is 10.15 µg/m^3.

4 (b) What county has the highest level of poverty in California?

- Drop counties with NA values for poverty
- Group by California county
- Take a weighted averaged based on total population of each census tract

```
data_pov <- data_clean %>%
  drop_na(poverty)

mean_pov <- data_pov %>%
  group_by(california_county) %>%
  summarize(weighted_mean = weighted.mean(poverty, total_population))

county <- subset(mean_pov, weighted_mean == max(mean_pov$weighted_mean))
county <- county[1]
max <- max(mean_pov$weighted_mean)
```

In California, Tulare is the county with the highest average poverty rate weighted by census tract total population with a rate of 50.147635 %.

5 (c) Make a histogram depicting the distribution of percent low birth weight and PM2.5

```
pm_lab <- expression(paste("PM2.5 (µg/m"^"3",")"))

p1 <- ggplot(data = data_clean) +
  geom_histogram(aes(x = low_birth_weight),
                 binwidth = 0.25) +
  theme_classic() +
  labs(y = "Frequency",
       x = "% of Low Birth Weights")

p2 <- ggplot(data = data_clean) +
  geom_histogram(aes(x = pm2_5),
                 binwidth = 0.5) +
  theme_classic() +
  labs(y = "Frequency",
       x = pm_lab)
```

Figure 1: Distributions of low birth weights and PM2.5 in California census tracts

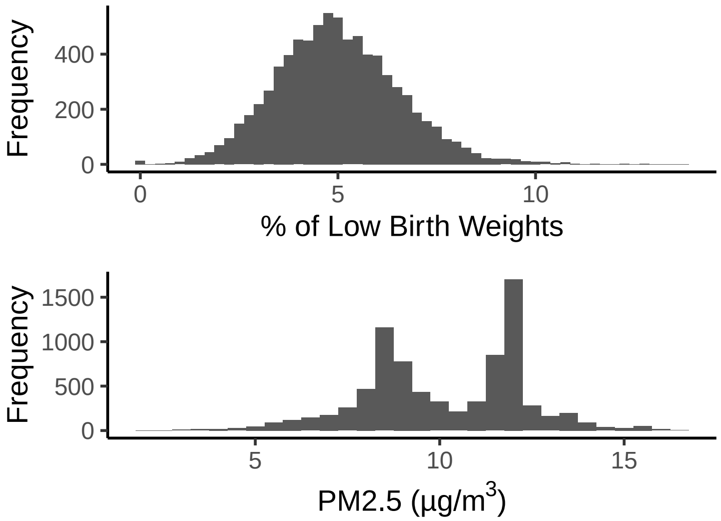

Figure 1 shows the distributions of % birth weights < 2500g and ambient concentrations of PM2.5 (µg/m^3) per census tract. Low birth weight data are approximately normally distributed while PM2.5 data is bimodal. Data is sourced from CalEnviroScreen 4.0.

6 (d) Estimate a OLS regression of LowBirthWeight on PM25. Report the estimated slope coefficient and its heteroskedasticity-robust standard error. Interpret the estimated slope coefficient.Is the effect of PM25 on LowBirthWeight statistically significant at the 5%?

To analyze the relationship between Low Birth Weight and PM2.5, we estimate the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{1}$$

where Y_i is Low Birth Weight for vehicle model i, X_{1i} is PM2.5 concentrations, X_{2i} is Poverty, and u_i the regression error term. We will consider a regression including only PM2.5, and a regression including PM2.5 and Poverty.

```r
mdl <- lm_robust(low_birth_weight ~ pm2_5, data=data_clean)

mdl_clean <- mdl %>%
  broom::tidy()
```

The estimated slope coefficient for PM2.5 (β_1) = 0.1179305, meaning that a 1 µg/m^3 change in PM2.5 on average increases the low birth rate by 0.1179305. At the 5% significance level, the effect of PM2.5 on low birth rate is statistically significant because the p-value < 0.05. The heteroskedasticity-robust standard error for PM2.5 (β_1) = 0.0084024.

7 (f) Add the variable Poverty as an explanatory variable to the regression in (d). Interpret the estimated coefficient on Poverty. What happens to the estimated coefficient on PM25, compared to the regression in (d). Explain.

```
mdl2 <- lm_robust(low_birth_weight ~ pm2_5 + poverty, data=data_clean)

mdl2_clean <- mdl2 %>%
  broom::tidy()
```

The estimated slope coefficient for Poverty $(\beta_2) = 0.0274353$, meaning that a 1% change in Poverty on average increases Low Birth Rate by 0.0274353. Compared to the prior model, the slope coefficient of PM2.5 (β_1) has decreased from 0.0591077to because Poverty helps explain some of the change in Low Birth Weight. In other words, the prior model suffered from omitted variables bias and caused us to overestimate the impact of PM2.5 alone on Low Birth Weight.

Table 1 shows the estimated coefficients from estimating equation (1).

Table 1: PM2.5 and Poverty associate with Low Birth Rate in California census tracts

	LBW	
	(1)	(2)
PM2.5	0.118***	0.059***
	(0.008)	(0.008)
Poverty		0.027***
		(0.001)
Observations	7,808	7,805
R^2	0.025	0.117
Note:	*p<0.1; **p<0.05; ***p<0.01	

8 (g) From the regression in (f), test the null hypothesis that the effect of PM2.5 is equal to the effect of Poverty

H_0: PM2.5 $=$ Poverty

H_A: PM2.5 $\neq$ Poverty

```
mdl3 <- linearHypothesis(model = mdl2,
                hypothesis.matrix = c("pm2_5 - poverty = 0"),
                white.adjust = "hc2")

p <- round(mdl3$`Pr(>Chisq)`[2], 6)
```

Since the p-value $= 0.000243 < 0.05$, we can reject the null hypothesis that the effect of PM2.5 is equal to the effect of Poverty on Low Birth Weight.