EDS241: Assignment 1

Clarissa Boyajian

01/20/2022

In this assignment we compare public health and socioeconomic data for all 8,035 census tracts within California. The data are from the California EnviroScreen 4.0 tool, a mapping and data tool produced by the California Office of Environmental Health Hazards Assessment (OEHHA).

1 Load and clean data

The following code loads and cleans the data.

```
raw_data <- read.csv(here::here("data", "CES4_0FINAL_results.csv"))

clean_data <- raw_data %>%
  clean_names() %>%
  select(c("census_tract", "total_population", "california_county",
           "low_birth_weight", "pm2_5", "poverty"))
```

2 Calculate basic statistics of data

2.1 Average PM2.5 concentration

The code chunk below calculates the average ambient concentration of PM2.5 across all census tracts in California.

```
average_pm2_5 <- mean(clean_data$pm2_5)
```

a. The average ambient concentration of PM2.5 throughout all census tracts in California is 10.1527 micrograms per cubic meter (mg/m^3).

2.2 Poverty level in California counties

The code chunk below calculated the California county with the highest level of poverty in two ways.

- First, we calculate the county with the highest percentage of its population living at least two times below the federal poverty level. This is done simply for each county by taking the mean of the percentage of poverty for all census tracks in that county.
- Second, we calculate the county with the largest number of people living two times below the federal poverty level. This is done by summing the total population of all census tracts within each county and finding the average percentage of poverty for each county. These calculations are then multiplied to find the total number of people in each county living in poverty.

```
# county with highest % poverty
county_average_poverty <- clean_data %>%
  group_by(california_county) %>%
  summarise(county_average_poverty = mean(poverty, na.rm = TRUE))

county_average_poverty_highest <-
  filter(county_average_poverty, county_average_poverty == max(county_average_poverty))

# county with most people in poverty
total_state_population <- sum(clean_data$total_population)

county_population_poverty <- clean_data %>%
  group_by(california_county) %>%
  summarise(total_county_population = sum(total_population, na.rm = TRUE),
            county_average_poverty_percent = mean(poverty, na.rm = TRUE),
            total_county_population_poverty =
              total_county_population * (county_average_poverty_percent / 100))

county_most_impoverished <-
  filter(county_population_poverty,
         total_county_population_poverty == max(total_county_population_poverty))
```

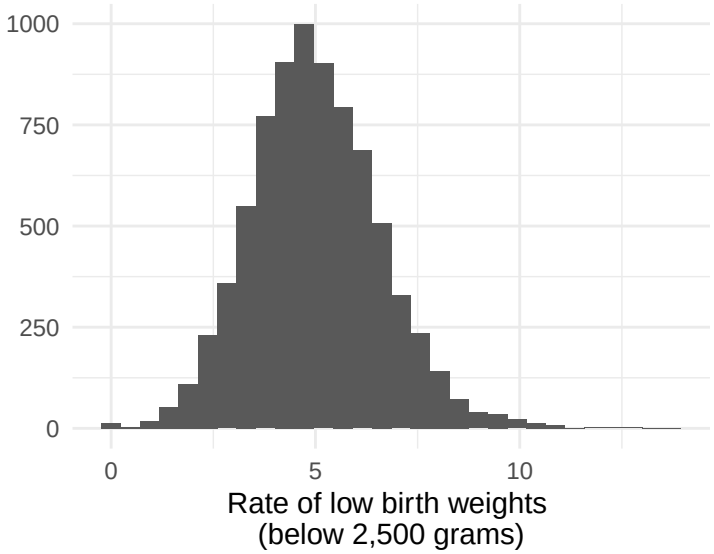
b. The county with the highest average percentage of poverty is Tulare County with 51.82% of the population living at least below two times the federal poverty level. The county with the highest number of people living in poverty is Los Angeles with 3541653 people living at least two times below the federal poverty level.

2.3 Low birth weight and PM2.5 distribution

The code chunk below shows how to produce a histogram of low birth weight percentages.

```
plot_low_birth_weight <- ggplot(clean_data, aes(x = low_birth_weight)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "Rate of low birth weights \n(below 2,500 grams)",
       y = "")
```

Figure 1: Histogram of CA Census Tract Low Birth Weight Percentages

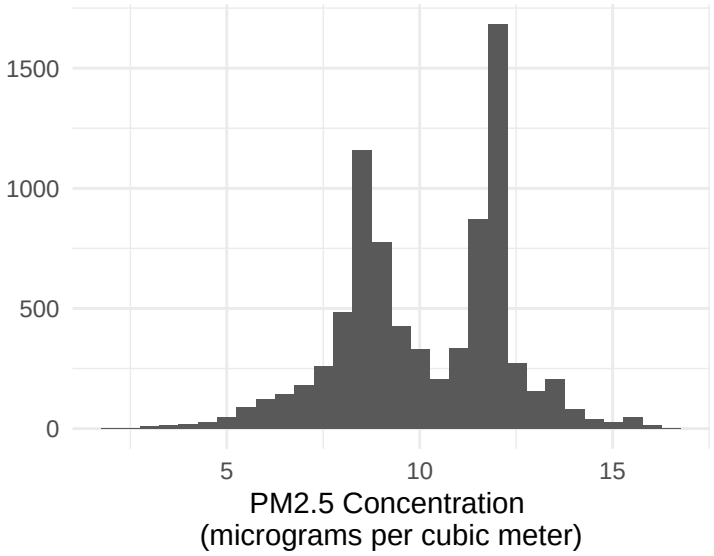


c. Figure 1 shows a histogram of the rate of births with weight less than 2,500 grams in all California census tracts. We can see that ~5% is the most common percentage of low birth weights throughout California census tracts.

The code chunk below shows how to produce a histogram of average ambient PM2.5 concentration levels.

```
plot_pm2_5 <- ggplot(clean_data, aes(x = pm2_5)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "PM2.5 Concentration \n(micrograms per cubic meter)",
       y = "")
```

Figure 2: Histogram of Average PM2.5 Concentration within CA Census Tracts



c. Figure 2 shows a histogram of the average ambient concentration of PM2.5 in all California census tracts. We can see that ~12 mg/m^3 is the most common concentration level with a second peak showing ~8.5 mg/m^3 as the second most common concentration level.

3 Run and interpret regression models

3.1 Impact of PM2.5 on birth weight

To analyze the relationship between the percentage of low birth rates and the average concentration of PM2.5 we estimate the following regression:

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

where Y_i is the rate of births with low weight (below 2,500g) for each census tract i, X_{1i} is the average PM2.5 concentration measured in micrograms per m^3, and u_i is the regression error term.

The code chunks below calculates the linear regression stated above in equation (1).

```
# estimate coefficients (inline reference)
model_d_robust <- lm_robust(formula = low_birth_weight ~ pm2_5, data = clean_data)

## for use in `stargazer` table:
# get coefficients
model_d <- lm(formula = low_birth_weight ~ pm2_5, data = clean_data)
# get robust standard error
se_model_d <- starprep(model_d)
```

```
stargazer(model_d, se = se_model_d,
          type = "latex", ci = FALSE, no.space = TRUE,
          header = FALSE, omit = c("Constant"),
          omit.stat = c("adj.rsq", "ser", "f"),
          covariate.labels = c("PM2.5"),
          dep.var.labels = c("Low Birth Weight"),
          dep.var.caption = c(""),
          notes = c("Robust standard errors paranthese."),
          title = "PM2.5 and Low Birth Weight",
          table.placement = "H")
```

Table 1: PM2.5 and Low Birth Weight

	Low Birth Weight
PM2.5	0.118***
	(0.008)
Observations	7,808
R^2	0.025
Note:	*p<0.1; **p<0.05; ***p<0.01
	Robust standard errors paranthese.

d. Table 1 shows the estimated slope coefficient ($\hat{\beta}_1$) is 0.1179 and the heteroskedasticity-robust standard error is 0.0084. We can interpret $\hat{\beta}_1$ to indicate that for each 1 mg/m^3 increase in PM2.5 concentration within a census tract, the percentage of births that are considered to be low weight increases by 11.79%. The effect of PM2.5 on low birth rates is statistically significant at the 1%.

3.2 Impact of PM2.5 and poverty on birth weight

To analyze the relationship between the rate of low birth rates, the average concentration of PM2.5, and poverty we estimate the following regression:

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

where Y_i is the percentage of births with low weight (below 2,500g) for each census tract i, X_{1i} is PM2.5 concentration, X_{2i} is average percentage of the population living at least two times below the federal poverty level, and u_i is the regression error term.

The code chunks below calculates the linear regression stated above in Equation 2.

```
# estimate coefficients (inline reference)
model_f_robust <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = clean_data)

## for use in `stargazer` table:
# get coefficients
model_f <- lm(formula = low_birth_weight ~ pm2_5 + poverty, data = clean_data)
# get robust standard error
se_model_f <- starprep(model_f)
```

```
stargazer(model_f, se = se_model_f,
          type = "latex", ci = FALSE, no.space = TRUE,
          header = FALSE, omit = c("Constant"),
          omit.stat = c("adj.rsq", "ser", "f"),
          covariate.labels = c("PM2.5", "Poverty"),
          dep.var.labels = c("Low Birth Rate"),
          dep.var.caption = c(""),
          notes = c("Robust standard errors paranthese."),
          title = "PM2.5, Poverty, and Low Birth Weight",
          table.placement = "H")
```

Table 2: PM2.5, Poverty, and Low Birth Weight

	Low Birth Rate
PM2.5	0.059***
	(0.008)
Poverty	0.027***
	(0.001)
Observations	7,805
R^2	0.117
Note:	*p<0.1; **p<0.05; ***p<0.01
	Robust standard errors paranthese.

f. Table 2 shows the estimated coefficient ($\hat{\beta}_2$) of poverty as 0.0274. This means that if you hold PM2.5 concentration constant, a 1% increase in poverty within a census tract will results in the rate of low birth weights to increase by 11.82%. The estimated coefficient on PM2.5 is roughly half of what it was in the previous regression (question d). $\hat{\beta}_2$ is now 0.0591 as compared to previously being 0.1179. This is unsurprising as there is likely a correlation between poverty and one's likelihood of living in a census tract with high PM2.5 concentration. Therefore, the first model did not predict the estimated coefficient of PM2.5 accurately because u_i was correlated with the PM2.5 regressor (aka the first least squared assumption was violated).

4 Null hypothesis test

The code chunk below tests the null hypothesis that the effect of PM2.5 concentration is equal to the effect of the rate of poverty on the rate of low birth weights within a given census tract.

```
model_hyp_test <- car::linearHypothesis(model = model_f_robust,
                                        c("pm2_5 = poverty"),
                                        white.adjust = "hc2")
pr_chisq <- model_hyp_test$`Pr(>Chisq)`[2]
```

g. Based on the joint hypothesis test above, we can reject the null hypothesis that the effects of PM2.5 concentration and poverty are equal because the p-value is 0.0002. This is statistically significant at 0.01%.