EDS241: Assignment 01 - CalEnviroScreen

Julia Parish

2022/01/21

This statistical analysis was completed as an assignment for the course, Environmental Data Science 241: Environmental Policy Evaluation. The data was sourced 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. **Source:** https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40

1 Clean and plot data

The following code loads and cleans the data. The variables selected were:

- **CensusTract**
- **TotalPopulation**
- **CaliforniaCounty**: the county where the census tract is located.
- **LowBirthWeight**: percent of census tract births with weight less than 2500g.
- **PM25**: ambient concentrations of PM2.5 in the census tract, in micrograms per cubic meters.
- **Poverty**: percent of population in the census tract living below twice the federal poverty line.

```
# Load data and convert to csv

convert("data/CES4.xlsx", "data/CES4.csv")

ces_raw <- import("data/CES4.csv")

# Clean data

ces_df <- ces_raw %>%
  clean_names() %>%
  select(census_tract,
         total_population,
         california_county,
         low_birth_weight,
         pm2_5,
         poverty) %>%
  mutate(low_birth_weight = round(as.numeric(low_birth_weight), 2)) %>%
  drop_na(low_birth_weight) %>%
  mutate(pm2_5 = round(as.numeric(pm2_5), 2)) %>%
  drop_na(poverty)

ces_df$california_county <- as.factor(ces_df$california_county)
```

2 Assignment Questions

2.1 A. What is the average concentration of PM2.5 across all census tracts in California?

```
pm2_5avg <- round(mean(ces_df$pm2_5), 3)
```

The average concentration of PM2.5 is 10.195 micrograms per cubic meter across all census tracts in California.

2.2 B. What county has the highest level of poverty in California?

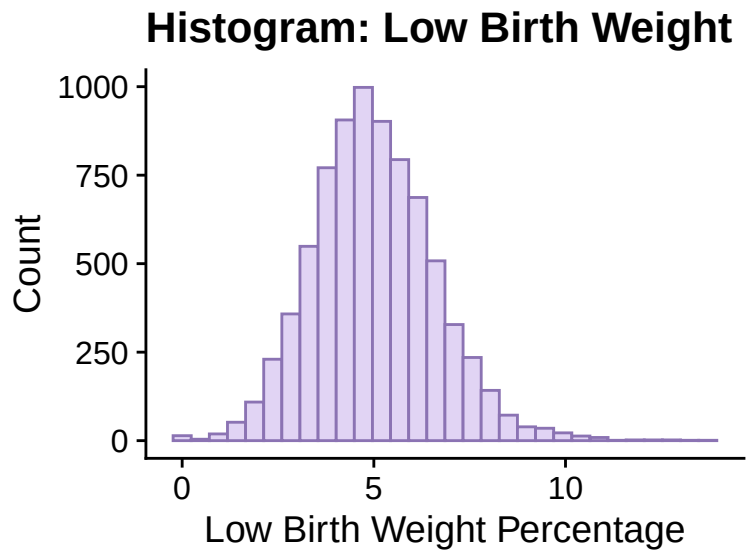
```
county_df <- ces_df %>%
  group_by(california_county) %>%
  summarize(weighted_mean = weighted.mean(poverty, total_population))

county <- subset(county_df, weighted_mean == max(county_df$weighted_mean))
county <- county[1]
povmax <- round(max(county_df$weighted_mean), 2)
```

Tulare County is the county with the highest average poverty rate weighted by total population in a census tract with a rate of 50.11 %.

2.3 C. Make a histogram depicting the distribution of percent low birth weight and PM2.5.

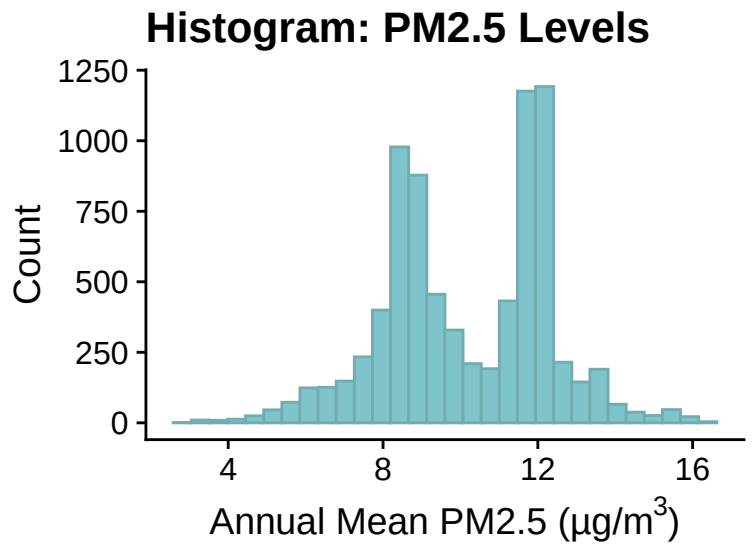
```
#Histogram for Birth Weight
ggplot(data = ces_df, aes(x = low_birth_weight)) +
  geom_histogram(color = "#8B73B3", fill = "#E1D4F4") +
  theme_cowplot(14) +
  labs(title = "Histogram: Low Birth Weight",
       x = "Low Birth Weight Percentage",
       y = "Count")
```



Low birth weight percentage has a normal distribution.

```r
pm_label <- expression(paste("Annual Mean PM2.5 (µg/m"^"3",")"))

#Histogram for PM2.5
ggplot(data = ces_df, aes(x = pm2_5)) +
  geom_histogram(color = "#71ACB1", fill = "#7FC4CA") +
  theme_cowplot(14) +
  labs(title = "Histogram: PM2.5 Levels",
       x = pm_label,
       y = "Count")
```



The PM2.5 concentrations has a bi-modal distribution

2.4 D. Estimate a OLS regression of LowBirthWeight on PM2.5. 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%?

$$Y_i = \beta_0 + \beta_1 PM2.5_{1i} + u_i \tag{1}$$

```r
lbw_model <- estimatr::lm_robust(low_birth_weight ~ pm2_5, data = ces_df)

huxtable::huxreg(lbw_model, error_pos = "right")
```

	(1)	
(Intercept)	3.799 ***	(0.089)
pm2_5	0.118 ***	(0.008)
N	7805	
R2	0.025	

*** p < 0.001; ** p < 0.01; * p < 0.05.

The effect of PM2.5 concentrations on low birth weight model produces the PM2.5 coefficient of 0.118. The heteroskedasticity-robust standard error is 0.008. **Beta1** equaling 0.118 represents the slope, and is interpreted as for every 1 unit increase in PM2.5 concentration, the birth weight percentage for a census tract increases by 0.118. This effect is statistically significant at the 5% level based on a p-value of less than 0.001.

2.5 F. Add the variable Poverty as an explanatory variable to the regression in Section 2.4 D. Interpret the estimated coefficient on Poverty. What happens to the estimated coefficient on PM2.5, compared to the regression in Section 2.4 D. Explain.

$$Y_i = \beta_0 + \beta_1 PM2.5_{1i} + \beta_2 Poverty_{2i} + u_i \tag{2}$$

```
pov_model <- estimatr::lm_robust(low_birth_weight ~ pm2_5 + poverty, data = ces_df)

huxtable::huxreg(pov_model, error_pos = "right")
```

	(1)	
(Intercept)	3.544 ***	(0.085)
pm2_5	0.059 ***	(0.008)
poverty	0.027 ***	(0.001)
N	7805	
R2	0.117	

*** p < 0.001; ** p < 0.01; * p < 0.05.

The robust multiple regression now shows the effect of PM2.5 concentrations on low birth weight produces has a coefficient of 0.118. The heteroskedasticity-robust standard error remains 0.008. The added variable, poverty, has a coefficient of 0.027. This is interpreted as, while holding PM2.5 constant, for every 1% increase in poverty for any census tract there is an expected 2.7% increase in low birth weight percentage. The PM2.5 coefficient has changed as poverty is another important explanatory variable on birth weight, which reduces the weight of influence on PM2.5 concentrations on low birth weight.

2.6 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 ≠ Poverty

```
lh_model <- linearHypothesis(model = pov_model, c("pm2_5=poverty"), white.adjust = "hc2")

p <- lh_model$`Pr(>Chisq)`[2]

p
```

```
## [1] 0.0002413898
```

The null hypothesis that the effect of PM2.5 is equal to the effect of poverty is rejected as the p-value of 0.0002414 «< 0.05.