EDS241: Assignment 1

Paloma Cartwright

2022-01-19

The data for this assignment come 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 .

The full data are contained in the file CES4.xls, which is available on Gauchospace (note that the Excel file has three "tabs" or "sheets"). The data is in the tab "CES4.0FINAL_results" and "Data Dictionary" contains the definition of the variables.

The following code loads and cleans the data.

```
# Load data and clean names
CES4_raw <- read_csv(here("ces4.csv")) %>%
  clean_names()


# Select columns wanted data
ces4 <- CES4_raw %>%
  select(census_tract, total_population,
         california_county, low_birth_weight, pm2_5, poverty)
```

The following code chunks answer the questions provided in the assignment.

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

```
# Avg PM2.5 Conc

avg_pm25_conc <- mean(ces4$pm2_5)
```

The average PM2.5 concentration across all census tracts in California is 10.1526999.

2 What county has the highest level of poverty in California?

```
# highest poverty

highest_poverty <- ces4 %>%
  group_by(california_county) %>%
  summarize(avg_poverty = na.omit(mean(poverty))) %>%
  arrange(desc(avg_poverty))
```

The county with the highest level of poverty is: Tulare. I calculated this by first grouping the data by county and then finding the average percentage of poverty per county. I chose the county with the highest mean percentage poverty as the answer to this question.

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

```
birth_weight_plot <- ggplot(data = ces4) +
  geom_histogram(aes(x = low_birth_weight)) +
  theme_minimal()+
  labs(y = "Frequency",
       x = "% of Birth Weight < 2500 grams")

pm2_5_plot <- ggplot(data = ces4) +
  geom_histogram(aes(x = pm2_5)) +
  theme_minimal() +
  labs(y = "Frequency",
       x = "PM 2.5 in micrograms per cubic meter")
```

```
birth_weight_plot / pm2_5_plot
```

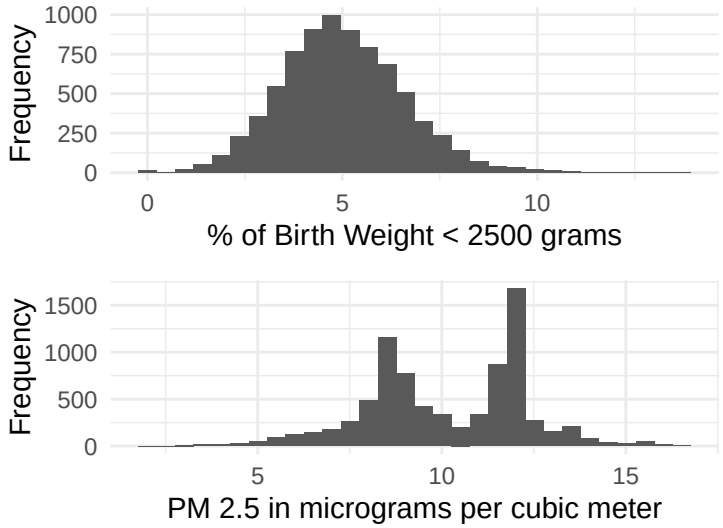

Figure 1: California Census Track Distributions of Low Birth Weight and PM2.5

4 Estimate a OLS regression of `low_birth_weight` on `pm2_5`. Report the estimated slope coefficient and its heteroskedasticity-robust standard error. Interpret the estimated slope coefficient. Is the effect of `pm2_5` on `low_birth_weight` statistically significant at the 5%?

```
model <- lm_robust(low_birth_weight ~ pm2_5, data = ces4)
summary(model)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5, data = ces4)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)   3.8010   0.088583   42.91 0.000e+00   3.6273   3.9746 7806
## pm2_5         0.1179   0.008402   14.04 3.256e-44   0.1015   0.1344 7806
##
## Multiple R-squared:  0.02499 ,   Adjusted R-squared:  0.02486
## F-statistic:   197 on 1 and 7806 DF,  p-value: < 2.2e-16
```

The slope coefficient for the regression of Low Birth Weight on PM2.5 is 0.1179305. The heteroskedacticity-robust standard error is 0.0084024. The slope coefficient tells us that for every 1 microgram per cubic meter increase in PM2.5 concentration, there is a 0.1179305 increase in the percentage of census tract births with weight less than 2500g. This is statistically significant at the 5% level.

4.0.1 Question e was removed

5 Add the variable `poverty` as an explanatory variable to the regression in (4). Interpret the estimated coefficient on Poverty. What happens to the estimated coefficient on PM25, compared to the regression in (4). Explain.

```
model2 <- lm_robust(low_birth_weight ~ pm2_5 + poverty, data = ces4)
summary(model2)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = ces4)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  3.54374   0.084733  41.823  0.000e+00  3.37764  3.70984 7802
## pm2_5        0.05911   0.008293   7.127  1.116e-12  0.04285  0.07536 7802
## poverty      0.02744   0.001002  27.374 1.287e-157  0.02547  0.02940 7802
```

```
##
## Multiple R-squared:  0.1169 ,    Adjusted R-squared:  0.1167
## F-statistic: 494.8 on 2 and 7802 DF,  p-value: < 2.2e-16
```

The estimated coefficient for poverty is 0.0274353. This tells us that holding PM2.5 constant, for every one percent increase in the population of the census tract living below twice the federal poverty line there is a 0.0274353 increase in the percentage of census tract births with weight less than 2500g. The coefficient for PM2.5 decreases in the second model which means that poverty explains some of the change in Low Birth Weights. The first model we examined had poverty contributing to omitted variable bias and overestimated the impact of PM2.5 on Low Birth Weight.

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

```
linhyp <- linearHypothesis(model2, c("pm2_5 = 0", "poverty = 0"),
                           white.adjust = "hc2")
linhyp
```

Res.Df	Df	Chisq	Pr(>Chisq)
7.8e+03			
7.8e+03	2	990	1.29e-215

With the p-value being $1.2877564 \times 10^{-215}$, we can reject the null hypothesis that the effect of PM2.5 is equal to the effect of poverty.