EDS241: Assignment 1

Halina Do-Linh

01/19/2022

For Assignment 1, we are exploring the variables `CensusTract`, `TotalPopulation`, `CaliforniaCounty`, `LowBirthWeight`, `PM25`, and `Poverty` from CalEnviroScreen 4.0, a mapping and data tool produced by the California Office of Environmental Health Hazards Assessment (OEHHA). The data comes from a variety of sources and covers all 8,035 census tracts in California.

Read in and Clean data

The following code loads and cleans the data.

```
# load data
ces <- read_csv(here::here("data", "ces4-final-results.csv"))

# clean data
ces_clean <- ces %>%
  janitor::clean_names() %>%
  select("census_tract",
         "total_population",
         "california_county",
         "low_birth_weight",
         "pm2_5",
         "poverty")
```

Question A

The code chunk below shows how to produce the average concentration of PM2.5 across all census tracts in California.

```
mean_pm25_all <- mean(ces_clean$pm2_5)
```

Answer: The average concentration of PM2.5 across all census tracts in California is 10.15.

Question B

The code chunk below shows how to produce the county with the highest level of poverty in California.

```
max_poverty_county <- ces_clean %>%
  filter(poverty != is.na(poverty)) %>%
  group_by(california_county) %>%
  summarize(weighted_mean_poverty = weighted.mean(poverty, total_population))
```

Answer: The county with the highest level of poverty in California is Tulare.

Question C

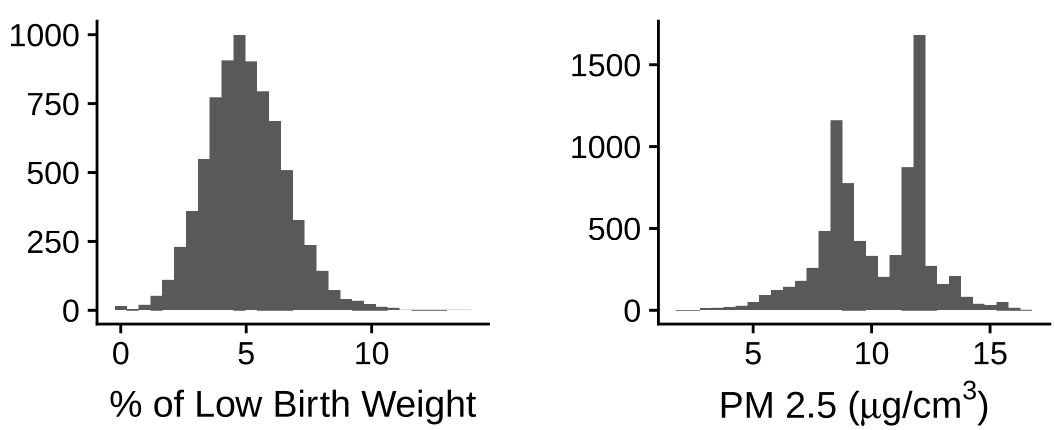
The code chunks below show how to produce histograms of `LowBirthWeight` and `PM25`.

```
# histogram low birth weight
low_birth_weight_hist <- ces_clean %>%
  ggplot(aes(x = low_birth_weight)) +
  geom_histogram() +
  theme_cowplot(14) +
  labs(x = "% of Low Birth Weight",
       y = "")
```

```
# histogram PM2.5
pm2_5_hist <- ces_clean %>%
  ggplot(aes(x = pm2_5)) +
  geom_histogram() +
  theme_cowplot(14) +
  labs(x = expression(paste("PM 2.5 ", "(", mu, "g/", cm^3, ")")),
       y = "")
```

Left: Distribution of Percentage of Low Birth Weight

Right: Distribution of PM 2.5 (μg/cm^3)



Question D

The code chunk below shows how to produce an OLS regression of `LowBirthWeight` on `PM25`.

$$lowbirthweight_i = \beta_0 + \beta_1 pm2.5_{1i} + u_i \tag{1}$$

```
model1 <- lm_robust(formula = low_birth_weight ~ pm2_5, data = ces_clean)

huxreg(model1)
```

	(1)
(Intercept)	3.801 ***
	(0.089)
pm2_5	0.118 ***
	(0.008)
N	7808
R2	0.025

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

Answer: The estimated β_1 slope coefficient is 0.118 and its heteroskedasticity-robust standard error is 0.008. The estimated β_1 slope coefficient tells us that for every 1 $\mu g/cm^3$ increase in PM2.5 we expect the percentage of low birth weights to increase by 0.118. The effect of `PM25` on `LowBirthWeight` is statistically significant at a significance level of 1% and therefore at 5% as well.

Question F

The code chunk below shows how to produce a multiple linear regression of `LowBirthWeight` with `PM25` and `poverty` as explanatory variables.

$$lowbirthweight_i = \beta_0 + \beta_1 pm2.5_{1i} + \beta_2 poverty_{2i} + u_i \tag{2}$$

```
model2 <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = ces_clean)

huxreg(model2)
```

	(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.

Answer: The estimated β_2 coefficient on `poverty` tells us that for every 1% increase in poverty we expect the percentage of low birth weight to increase by 0.027. The estimated β_1 coefficient on PM 2.5 went down by half compared to the estimated β_1 coefficient in `model1` from Question D. This is due to omitted variables bias. And in this case, `model1` from Question D was exaggerating the effect of PM 2.5 since `model2` shows that `poverty` also has a significant effect on `LowBirthWeight`.

Question G

The code chunk below is using hypothesis testing to test the null hypothesis that the effect of `PM25` is equal to the effect of `poverty` based on `model2`.

```
# test that pm2_5 = poverty or pm2_5 - poverty = 0
linearHypothesis(model2, c("pm2_5=poverty"), white.adjust = "hc2")
```

Res.Df	Df	Chisq	Pr(>Chisq)
7.8e+03			
7.8e+03	1	13.5	0.000243

Answer: We reject the null hypothesis that the effect of `PM25` is equal to the effect of `poverty` because the p-value is statistically significant at the 1% significance level.