

EDS241: Assignment 1

Joe DeCesaro

01/20/2022

1 Assignment 1

For the assignment, you will need the following variables: **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), and **Poverty** (percent of population in the census tract living below twice the federal poverty line).

1.1 Clean data

The following code loads and cleans the data.

```
# Read in the first sheet and clean up
data_sheet1 <- read_xlsx(here("CES4.xlsx"),
                        sheet = 1,
                        na = "NA") %>%

clean_names() %>%
select(census_tract, total_population, california_county, low_birth_weight, pm2_5, poverty)
```

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

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

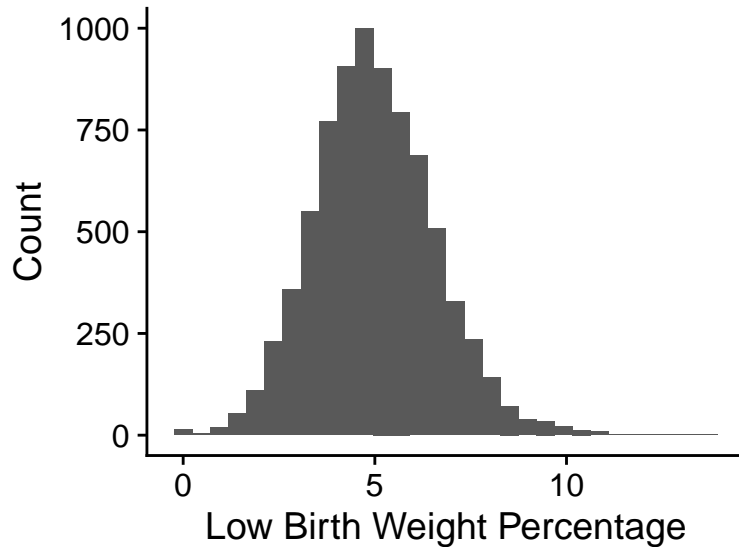
1.3 b) What county has the highest level of poverty in California?

The county with the highest average level of poverty in California is Tulare County

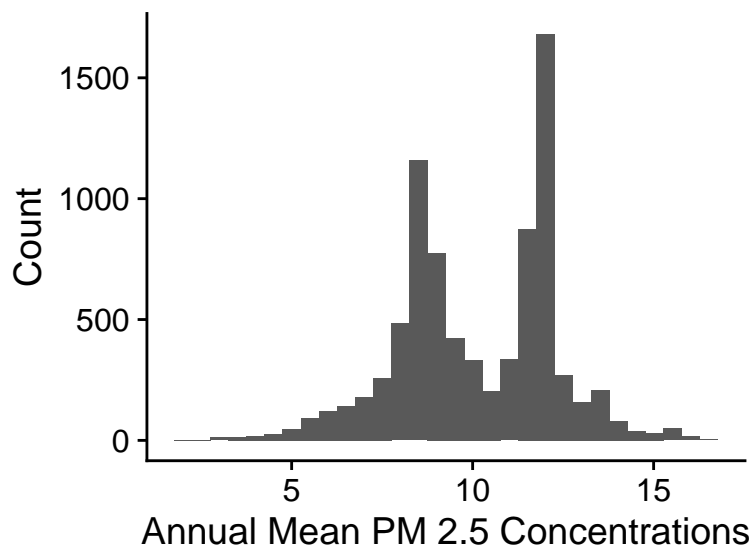
```
pov <- data_sheet1 %>%
  group_by(california_county) %>%
  summarise(mean(poverty))
```

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

```
ggplot(data = data_sheet1, aes(x = low_birth_weight)) +  
  geom_histogram() +  
  theme_cowplot(14) +  
  labs(x = "Low Birth Weight Percentage",  
       y = "Count")
```



```
ggplot(data = data_sheet1, aes(x = pm2_5)) +  
  geom_histogram() +  
  theme_cowplot(14) +  
  labs(x = "Annual Mean PM 2.5 Concentrations",  
       y = "Count")
```



- 1.5 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%?

```
model_1 <- estimatr::lm_robust(low_birth_weight ~ pm2_5, data = data_sheet1)

tidy_model_1 = broom::tidy(model_1)

tidy_model_1 %>%
  dplyr::filter(term != '(Intercept)') %>%
  dplyr::select(term, estimate, std.error, p.value) %>%
  knitr::kable() %>%
  kableExtra::kable_styling(full_width = FALSE,
                             latex_options = "HOLD_position")
```

term	estimate	std.error	p.value
pm2_5	0.1179305	0.0084024	0

The estimated slope coefficient is 0.118 and its heteroskedasticity-robust standard error is 0.008. The slope coefficient can be interpreted that for every 1 unit increase in PM2.5 we can expect a low birth weight percentage for the census tract to increase by 0.118. As the standard error is within the bounds of the confidence interval it is statistically significant.

- 1.6 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 PM2.5, compared to the regression in (d). Explain.

```
model_2 <- estimatr::lm_robust(low_birth_weight ~ pm2_5 + poverty, data = data_sheet1)

tidy_model_2 = broom::tidy(model_2)

tidy_model_2 %>%
  dplyr::filter(term != '(Intercept)') %>%
  dplyr::select(term, estimate, std.error, p.value) %>%
  knitr::kable() %>%
  kableExtra::kable_styling(full_width = FALSE,
                             latex_options = "HOLD_position")
```

term	estimate	std.error	p.value
pm2_5	0.0591077	0.0082932	0
poverty	0.0274353	0.0010022	0

The estimated coefficient of poverty in this model is 0.027. This can be interpreted to mean that for every 1% increase in “poverty” there is an expected 0.027 increase in low birth weight percentage for the census tract while holding PM2.5 constant. The estimated coefficient for PM2.5 decreases from 0.118 to 0.059 in this model compared to the previous. This happens because in the previous model the PM2.5 was trying to account for all of the variation in low birth weight percentages and now that change is, so to speak, “divided” with the poverty metric.

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

```
lin_model <- linearHypothesis(model = model_2, c("pm2_5=poverty"), white.adjust = "hc2")

tidy_lin_model = broom::tidy(lin_model)

tidy_lin_model %>%
  dplyr::select(statistic, p.value) %>%
  dplyr::filter(statistic > 0) %>%
  knitr::kable() %>%
  kableExtra::kable_styling(full_width = FALSE,
                             latex_options = "HOLD_position")
```

statistic	p.value
13.46823	0.0002426

We can reject the null hypothesis that the effect of PM2.5 is equal to the effect of Poverty as the p-value is statistically significant.