EDS241: Assignment 1

Wylie Hampson

01/21/2022

In this assignment, we use data from CalEnviroScreen 4.0, a mapping 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.

1 Clean and plot data

The following code loads and cleans the data, then selects the columns that we are interested in.

```
# Load data

data <- read_excel(here("data", "CES4.xlsx"), na = "NA")

# Clean data

data <- data %>% clean_names()

# Select the columns that we are interested in.

data <- data %>%
  select(
    census_tract,
    total_population,
    california_county,
    low_birth_weight,
    pm2_5,
    poverty
  )
```

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

```
mean_pm <- round(mean(data$pm2_5), 3)
mean_pm
```

```
## [1] 10.153
```

The mean PM2.5 concentration across all census tracts in California is **10.153** $\mu g/m^3$.

Question b: What county has the highest level of poverty in California?

```
poverty_by_county <- data %>%
  group_by(california_county) %>%
  summarize(county_poverty = mean(poverty, na.rm = TRUE)) %>%
  arrange(desc(county_poverty)) %>%
  head()

highest_poverty_rate <- round(max(poverty_by_county$county_poverty), 1)
highest_poverty_county <- poverty_by_county$california_county[1]

poverty_by_county
```

california_county	county_poverty
Tulare	51.8
Del Norte	48.4
Imperial	47.8
Merced	47.3
Kern	47.2
Fresno	45.8

In the above summary table we can see that the California county that has the highest rate of poverty is Tulare with a poverty rate of 51.8%. But which county has the most people living in poverty?

```
people_in_poverty <- data %>%
  group_by(california_county) %>%
  summarize(county_poverty = mean(poverty, na.rm = TRUE),
            county_pop = sum(total_population),
            people_in_poverty = (county_poverty / 100) * county_pop) %>%
  arrange(desc(people_in_poverty)) %>%
  head()

people_in_poverty
```

california_county	county_poverty	county_pop	people_in_poverty
Los Angeles	35.1	1.01e+07	3.54e+06
San Diego	28.1	3.32e+06	9.31e+05
Riverside	35	2.41e+06	8.43e+05
San Bernardino	38.2	2.15e+06	8.21e+05
Orange	24.4	3.17e+06	7.75e+05
Sacramento	33.2	1.52e+06	5.07e+05

```
most_poverty_county <- people_in_poverty$california_county[1]
most_people_county <- round(people_in_poverty$people_in_poverty[1], 0)
```

In this summary table we can see that the California county that has the highest number of people that are living in poverty is Los Angeles County, with 3541653 people living in poverty.

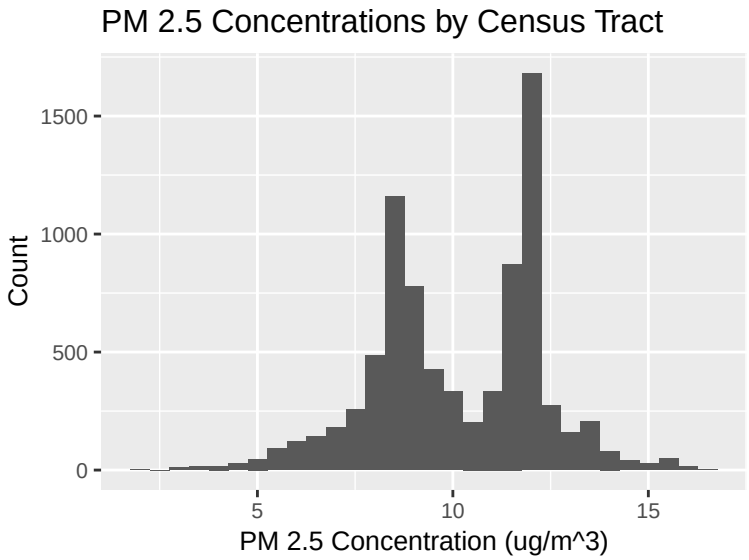
Question c: Make a histogram depicting the distribution of percent low birth weight and PM2.5.

```
birth_hist <- ggplot(data = data, aes(x = low_birth_weight)) +
  geom_histogram() +
  labs(x = "Percent of Births With Weight Less Than 2500g",
       y = "Count",
       title = "Percent of Birth Weights Below 2500g by Census Tract") +
  theme(text = element_text(size = 10),
        title = element_text(size = 7))

birth_hist
```


```
pm_hist <- ggplot(data = data, aes(x = pm2_5)) +
  geom_histogram() +
  labs(x = "PM 2.5 Concentration (ug/m^3)",
       y = "Count",
       title = "PM 2.5 Concentrations by Census Tract") +
  theme(text = element_text(size = 10))

pm_hist
```



Question d: 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 PM25 on LowBirthWeight statistically significant at the 5%?

```
model_1 <- lm_robust(formula = low_birth_weight ~ pm2_5, data = data)
summary(model_1)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5, data = data)
##
## 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
```

```
model_1_table <- broom::tidy(model_1) %>%
  dplyr::select(term, estimate, std.error, p.value) %>%
  knitr::kable()

model_1_table
```

term	estimate	std.error	p.value
(Intercept)	3.8009877	0.0885829	0
pm2_5	0.1179305	0.0084024	0

Here the estimated slope coefficient is 0.1179, meaning that with every unit increase of PM 2.5 concentration, we would expect to see a 0.1179% increase of babies that are born under 2500 grams, on average. The

heteroskedasticity-robust standard error is 0.0084 with standard error type of HC2. The p-value here is extremely close to 0, so we can reject the null hypothesis that PM 2.5 concentration does not effect low birth weight percentages.

Question 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 <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = data)

model_2_table <- broom::tidy(model_2) %>%
  dplyr::select(term, estimate, std.error, p.value) %>%
  knitr::kable()

model_2_table
```

term	estimate	std.error	p.value
(Intercept)	3.5437420	0.0847329	0
pm2_5	0.0591077	0.0082932	0
poverty	0.0274353	0.0010022	0

Here the estimated coefficient on poverty is 0.0274, meaning that if pm2_5 is held constant, we would expect to see a 0.0274% increase of babies that are born under 2500 grams with every 1 percent increase of people living in poverty, on average. In this model the estimated coefficient for PM 2.5 is now 0.0591. It makes sense that the estimated coefficient for PM 2.5 got smaller in model 2, because in model 1 poverty was a missing variable that does help explain low birth weight percentages. So by including that missing variable, it would reduce the slope coefficient of PM 2.5.

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

```
h_test <- linearHypothesis(model_2, c("pm2_5 = poverty"), white.adjust = "hc2") %>%
  broom::tidy() %>%
  knitr::kable()

h_test
```

res.df	df	statistic	p.value
7803	NA	NA	NA
7802	1	13.46823	0.0002426

From the hypothesis test above we can see that we get a p-value back of about 0.00024. This is a very low p-value, so at the 5% significance level, we can reject the null hypothesis that the effect of PM 2.5 is equal to the effect of poverty on low birth weight percentage.