EDS241: Assignment 1

Marie Rivers

01/21/2022

In this assignment, we use air quality data in R to investigate the relationship between PM2.5 and low birth weight in California. The data came 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 Read and Clean Data

```
data <- read_xlsx(here("data", "CES4.xlsx"), sheet = "CES4.0FINAL_results") %>%
  rename(CensusTract = "Census Tract", TotalPopulation = "Total Population", CaliforniaCounty = "Califo
  select(CensusTract, TotalPopulation, CaliforniaCounty, LowBirthWeight, PM25, Poverty) %>%
  mutate(LowBirthWeight = as.numeric(LowBirthWeight))
```

2 Question a:

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

```
mean_pm25 <- mean(data$PM25)
```

Answer: The average concentration of PM2.5 across all census tracts in California is 10.15 $\mu g/m^3$.

3 Question b:

What county has the highest level of poverty in California?

```
county_pov <- data %>%
  mutate(pov_per_capita = Poverty/TotalPopulation) %>%
  group_by(CaliforniaCounty) %>%
  summarise(mean_pov = mean(Poverty, na.rm = TRUE),
            mean_pov_per_capita = mean(pov_per_capita, na.rm = TRUE))

county_max_pov <- county_pov$CaliforniaCounty[which.max(county_pov$mean_pov)]
county_max_pov_per_capita <- county_pov$CaliforniaCounty[which.max(county_pov$mean_pov_per_capita)]
```

Answer: Tulare County has the highest level of poverty in California based on mean poverty for all census tracts in each county. I decided not to use the county with the highest level of poverty per capital (which would have been Alpine) because according to the CES4 data dictionary the `Poverty` variable represents the percent of population living below two times the federal poverty level.

Note: there are 75 census tracts with NA values for poverty

4 Question c:

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

Figure 1: Percent Low Birth Weight for California Census Tracts

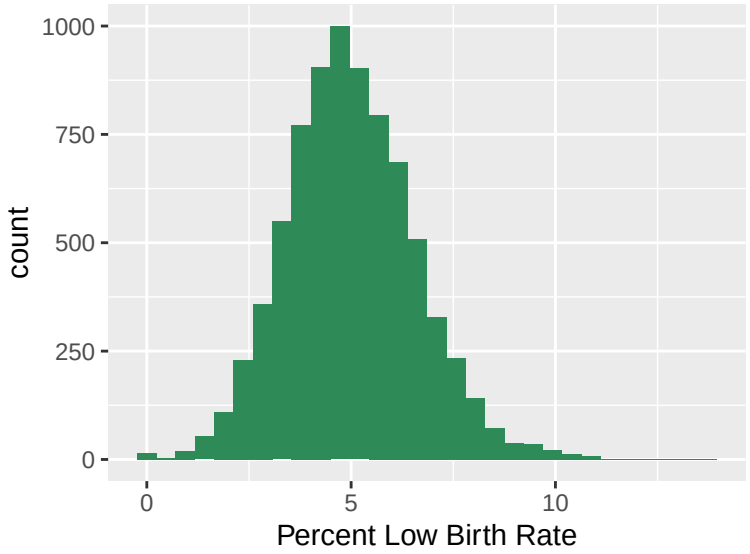

Figure 1 shows the distribution of percent low birth weights for California Census Tracts as reported by the California Office of Environmental Health Assessment (OEHHA). Data Source: CalEnviroScreen 4.0.

Figure 2: Air Quality for California Census Tracts

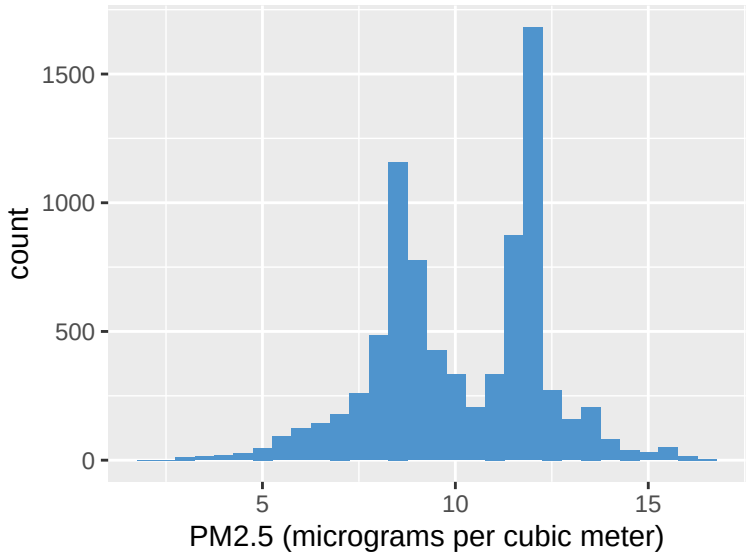

Figure 2 shows the distribution of annual mean PM 2.5 concentrations for California Census Tracts as reported by the California Office of Environmental Health Assessment (OEHHA). Data Source: CalEnviroScreen 4.0.

5 Question d:

Estimate a OLS regression of `LowBirthWeight` on `PM25`. 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% level?

$$\text{percent low birth weight}_i = \beta_0 + \beta_1 \cdot PM2.5 + \varepsilon_i$$

```
ols_model_robust <- lm_robust(formula = LowBirthWeight ~ PM25, data = data)
```

```
# use lm to estimate coefficients
ols_model <- lm(formula = LowBirthWeight ~ PM25, data = data)

# adjust standard errors using estimatr::starprep() instead of estimatr::lm_robust()
se_ols_model <- starprep(ols_model)
```

Table 1: Low Birth Weight and PM 2.5 concentration

	percent low birth weight
PM 2.5 concentration	0.118***
	(0.008)
Observations	7,808
R^2	0.025
Note:	*p<0.1; **p<0.05; ***p<0.01
	Robust standard errors in parentheses

```r
intercept <- ols_model_robust$coefficients[1]
intercept
```

```
## (Intercept)
##    3.800988
```

```r
slope_coef <- ols_model_robust$coefficients[2]
slope_coef
```

```
##      PM25
## 0.1179305
```

```r
std_err <- ols_model_robust$std.error[2]
std_err
```

```
##        PM25
## 0.008402393
```

```r
ci_lower <- ols_model_robust$conf.low[2]
ci_lower
```

```
##      PM25
## 0.1014596
```

```r
ci_upper <- ols_model_robust$conf.high[2]
ci_upper
```

```
##      PM25
## 0.1344015
```

Answer: The estimated slope coefficient for this model is 0.118. The estimated slope coefficient means that for a one unit increase in the concentration of PM2.5, percent low birth weight will increase by 0.118. The heteroskedasticity-robust standard error for this estimated slope coefficient is 0.008. Since the p-value is $<$ 0.05, the effect of PM2.5 on percent low birth weight is statistically significant at the 5 % level.

```r
ols_model_plot <- ggplot(data = data, aes(x = PM25, y = LowBirthWeight)) +
  geom_point() +
  geom_smooth(method = lm)
```

Figure 3: OLS Regression of Low Birth Weight on PM 2.5

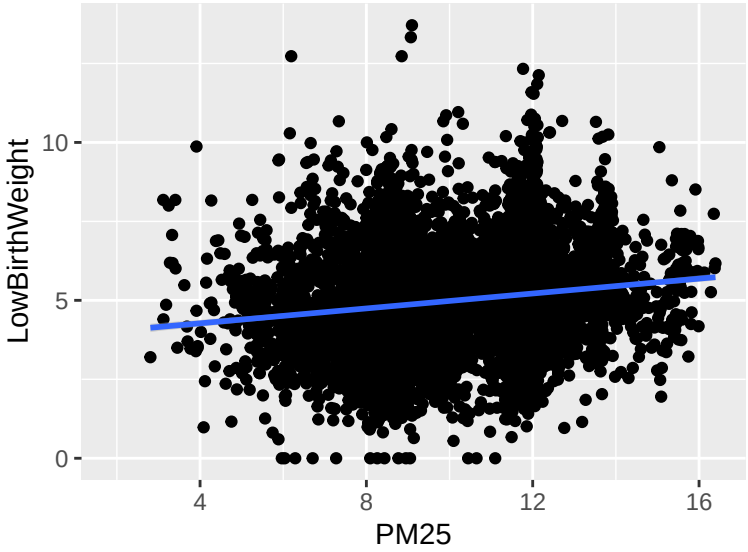

Figure 3 shows a positive correlation between low birth weight and PM 2.5 based on an Ordinary Least Squares Regression model of low birth weight on PM 2.5.

6 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.

```
pm25_poverty_model_robust <- lm_robust(formula = LowBirthWeight ~ PM25 + Poverty, data = data)
```

```
# use lm to estimate coefficients
pm25_poverty_model <- lm(formula = LowBirthWeight ~ PM25 + Poverty, data = data)

# adjust standard errors using estimatr::starprep() instead of estimatr::lm_robust
se_pm25_poverty_model <- starprep(pm25_poverty_model)
```

Table 2: Low Birth Weight on PM 2.5 concentration and Poverty

	percent low birth weight	
	(1)	(2)
PM 2.5 concentration	0.118***	0.059***
	(0.008)	(0.008)
Poverty		0.027***
		(0.001)
Observations	7,808	7,805
R^2	0.025	0.117
Note:		*p<0.1; **p<0.05; ***p<0.01
		Robust standard errors in parentheses

```
poverty_coef <- pm25_poverty_model_robust$coefficients[3]
poverty_coef
```

```
##    Poverty
## 0.02743528
```

```
pm25_coef_2 <- pm25_poverty_model_robust$coefficients[2]
pm25_coef_2
```

```
##       PM25
## 0.05910773
```

The estimated coefficient on poverty is 0.027. This mean that for a one unit increase in poverty, percent low birth weight will increase by 0.027 while holding PM2.5 constant. In this model, the estimated coefficient on PM2.5 is lower than the coefficient estimated with the model that did not include poverty. This suggests that PM2.5 is not the only variable to influences low birth weight.

7 Question g:

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

null hypothesis: There is no difference in the estimated slope coefficients for PM2.5 and poverty.

$$H_0 : \beta_{1,PM2.5} - \beta_{2,poverty} = 0$$

alternative hypothesis: There is a difference in the estimated slope coefficient for PM2.5 and poverty.

$$H_A : \beta_{1,PM2.5} - \beta_{2,poverty} \neq 0$$

```
linearHypothesis(model = pm25_poverty_model, hypothesis.matrix = c("PM25=Poverty"), white.adjust = "hc2
```

```
## Linear hypothesis test
##
## Hypothesis:
## PM25 - Poverty = 0
##
## Model 1: restricted model
## Model 2: LowBirthWeight ~ PM25 + Poverty
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   7803
## 2   7802  1 13.468 0.0002443 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer: Based on these results, we reject the null hypothesis that $\beta_{1,PM2.5} = \beta_{2,poverty}$.