EDS241: Assignment 1

Desik Somasundaram

01/21/2022

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

1 Clean and plot data

The following code loads and cleans the data.

```r
# Load data
CESdata <- read_excel("CES4.xlsx", sheet = 1, na = "NA")


# Clean data
CESdata <-janitor::clean_names(CESdata)
```

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

```r
(avgPM25 <- round(mean(CESdata$pm2_5),3))
```

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

The average concentration of PM2.5 across all census tracts in California is ***10.153 micrograms per cubic meters***.

What county has the highest level of poverty in California?

```r
CESdata_summary <- CESdata %>%
                   group_by(california_county) %>%
                   mutate(wm_poverty = weighted.mean(poverty, total_population))
highestpovertycounty <- CESdata_summary[which.max(CESdata_summary$wm_poverty),3]
```

Using a weighted mean apporach that considers poverty and total population, the county has the highest level of poverty in California is ***Tulare***.

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

```
# Histogram

lowbwhist <- ggplot(CESdata, aes(x=CESdata$low_birth_weight))+
  geom_histogram()+
  labs(x = "% Low Birth Weight", y = "Count")

PM25hist <- ggplot(CESdata, aes(x=CESdata$pm2_5))+
  geom_histogram()+
  labs(x = "Ambient PM2.5 Level", y = "Count")
```

Figure 1: Percent Low Birth Weight Distribution in CA Census Tracts

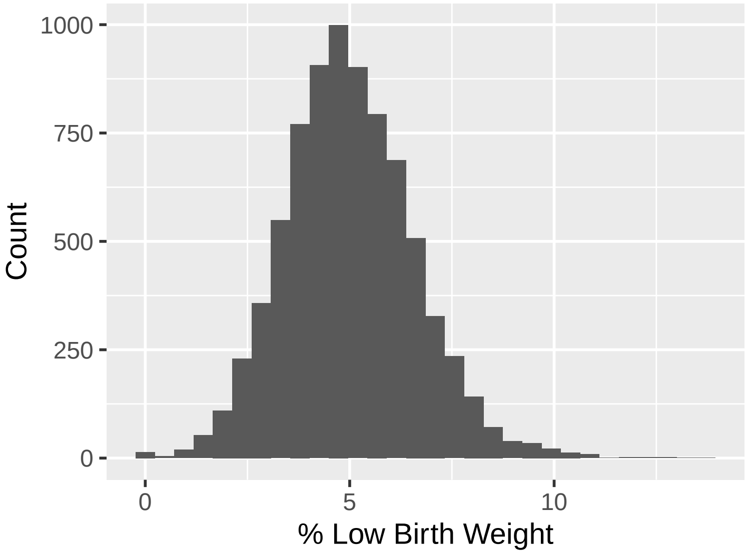

Figure 2: Ambient PM2.5 Level Distribution in CA Census Tracts

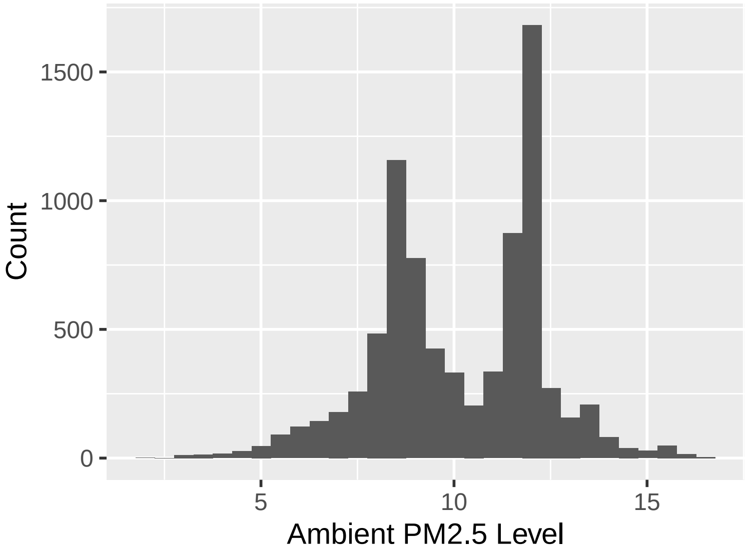

Figure 1 shows an approximately normal distribution for percent low birth weights while Figure 2 shows an approximately bimodal normal distribution for ambient PM2.5 levels.

2 Run and interpret regression models

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 PM2.5 on LowBirthWeight statistically significant at the 5%?

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \tag{1}$$

where Y_i is LowBirthWeight i, X_{1i} is the PM2.5 level, and u_i the regression error term. We will consider a regression including only PM2.5, and a regression including PM2.5 and Poverty.

In R, we run the following code:

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

Table 1 shows the estimated coefficients from estimating equation (1).

In model (1), the estimated β_1 coefficient implies that a 1 microgram per cubic meter increase in pm2_5 increases percent of census tract births with weight less than 2500g by 0.118. The effect of PM2.5 on LowBirthWeight is ***statistically significant at the 5%***.

) 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 PM25, compared to the regression in (d). Explain.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{2}$$

	(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.

here Y_i is LowBirthWeight i, X_{1i} is the PM2.5 level, X_{2i} is the poverty level and u_i the regression error term.

In R, we run the following code:

```r
model_2 <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data=CESdata)
```

Table 2 shows the estimated coefficients from estimating equation (2).

	(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.

In model (2), the estimated β_1 coefficient implies that a 1 microgram per cubic meter increase in pm2_5 increases percent of census tract births with weight less than 2500g by 0.059. the estimated β_2 coefficient implies that a 1 percent increase in poverty rate within a census tract increases percent of census tract births with weight less than 2500g by 0.027. The effect of PM2.5 on LowBirthWeight is still ***statistically significant at the 5%*** and the effect of Poverty on LowBirthWeight is also ***statistically significant at the 5%***. Adding the Poverty in model (2) reduces $\hat{\beta}_1$ from 0.118 to 0.059. This is likely due to omitted variable bias in model (1) which that more heavily weighs the effect of PM2.5 in the absence of other important variables such as Poverty.

Table 3 shows results from the linear hypothesis test whether the effect of PM2.5 and Poverty on LowBirthWeight are equivalent.

```
linearHypothesis(model_2,c("pm2_5 = poverty"), white.adjust = "hc2")
```

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

Based on the p-value, we reject the null that the effect of PM2.5 and Poverty are equal.