EDS241: Assignment 1

Joe DeCesaro

01/18/2022

1 Assignment 1

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

The full data are contained in the file CES4.xls, which is available on Gauchospace (note that the Excel file has three "tabs" or "sheets"). The data is in the tab "CES4.0FINAL_results" and "Data Dictionary" contains the definition of the variables.

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

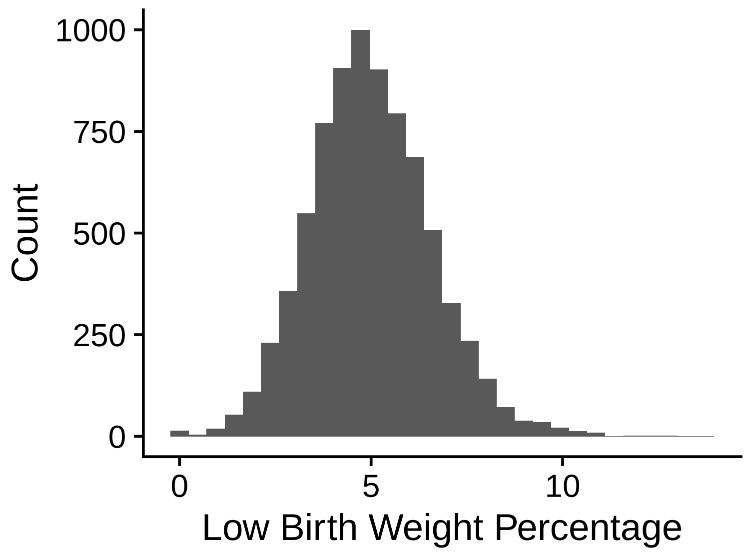
1.3 b) What county has the highest level of poverty in California?

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

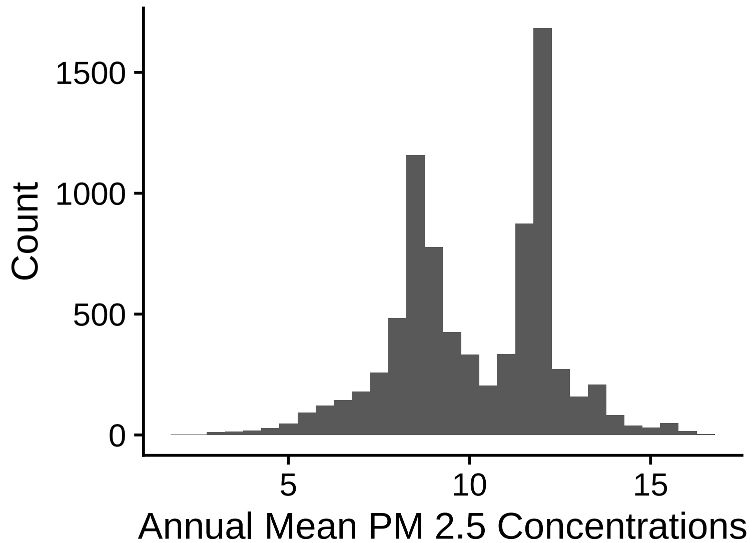
```r
# pov <- data_sheet1 %>%
#   group_by(california_county) %>%
#   summarise(pov_total = sum(total_population),
#             avg)

# Olivier will go over or need to clarify
```

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

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




```r
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 PM25. 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)
summary(model_1)
```

```
##
## Call:
## estimatr::lm_robust(formula = low_birth_weight ~ pm2_5, data = data_sheet1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   3.8010   0.088583   42.91
## pm2_5         0.1179   0.008402   14.04
##                                                 Pr(>|t|) CI Lower CI Upper
## (Intercept) 0.000000000000000000000000000000000000000000   3.6273   3.9746
## pm2_5       0.000000000000000000000000000000000000000003256   0.1015   0.1344
##               DF
## (Intercept) 7806
## pm2_5       7806
##
## Multiple R-squared:  0.02499 ,   Adjusted R-squared:  0.02486
## F-statistic:   197 on 1 and 7806 DF,  p-value: < 0.00000000000000022
```

The estimated slope coefficient is 0.118 and its heteroskedasticity-robust standard error is 0.008. The slope coefficient can be interpreted as 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 regressionin (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)
summary(model_2)
```

```
##
## Call:
## estimatr::lm_robust(formula = low_birth_weight ~ pm2_5 + poverty,
##     data = data_sheet1)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  3.54374   0.084733  41.823
## pm2_5        0.05911   0.008293   7.127
## poverty      0.02744   0.001002  27.374
##
## (Intercept) 0.0000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## pm2_5       0.0000000000111554859094103077673693205259373130108259408732607198544428683817386627197
## poverty     0.0000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##             CI Lower CI Upper   DF
## (Intercept)  3.37764  3.70984 7802
## pm2_5        0.04285  0.07536 7802
## poverty      0.02547  0.02940 7802
##
## Multiple R-squared:  0.1169 ,    Adjusted R-squared:  0.1167
## F-statistic: 494.8 on 2 and 7802 DF,  p-value: < 0.00000000000000022
```

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 0.043 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 changes 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

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

```
## Linear hypothesis test
##
## Hypothesis:
## pm2_5 - poverty = 0
##
## Model 1: restricted model
```

```
## Model 2: low_birth_weight ~ pm2_5 + poverty
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   7803
## 2   7802  1 13.468  0.0002426 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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.