assignment_1

Juliet Cohen

1/13/2022

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).

```r
library(here)
library(readxl)
library(janitor)
library(tidyverse)
library(estimatr)
library(car)
```

```r
#data_xlsx <- read_excel(here("CES4.xlsx"))

#class(data_xlsx$"Low Birth Weight")

#data_xlsx <- data_xlsx %>%
#  unlist("Low Birth Weight") %>%
#  as.numeric("Low Birth Weight")

#relevant_data_xlsx <- data_xlsx %>%
#  select(census_tract, pm2_5, total_population, california_county, low_birth_weight, poverty)

#map(relevant_data_xlsx, ~sum(is.na(.)))

# import data as a csv so the class is normal but there is fewer data points
data <- read.csv(here("CES4_copy.csv")) %>%
  clean_names()

#colnames(data)
```

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

```r
# subset data for only relevant columns for this assignment
relevant_data <- data %>%
  select(census_tract, pm2_5, total_population, california_county, low_birth_weight, poverty)

# check if there are any NA values
map(relevant_data, ~sum(is.na(.)))
```

```
## $census_tract
## [1] 0
##
## $pm2_5
## [1] 0
##
## $total_population
## [1] 0
##
## $california_county
## [1] 0
##
## $low_birth_weight
## [1] 227
##
## $poverty
## [1] 75
```

```
# there are 75 NA values in the poverty col, and 227 NA values in LBW

# remove na rows
relevant_data <- na.omit(relevant_data)

avg_pm2_5 <- mean(relevant_data$pm2_5)
avg_pm2_5
```

```
## [1] 10.19529
```

The average ambient PM2.5 concentration across all census tracts in California is 10.1952898 micrograms per cubic meter

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

```
# get the mean poverty value for each county
mean_pov_county_df <- relevant_data %>%
  group_by(california_county) %>%
  summarise(mean_pov = mean(poverty), na.rm = TRUE)

# remove rows with NA
mean_pov_county_no_na <-  na.omit(mean_pov_county_df)

# find county with max avg poverty
#max_pov_county <- max(mean_pov_county_no_na$mean_pov)
#max_pov_county

# reduce dataframe to just row with max value
mean_pov_county <- mean_pov_county_no_na[which.max(mean_pov_county_no_na$mean_pov),]
mean_pov_county
```

```
## # A tibble: 1 x 3
##   california_county mean_pov na.rm
##   <chr>                <dbl> <lgl>
## 1 "Tulare "             51.5 TRUE
```

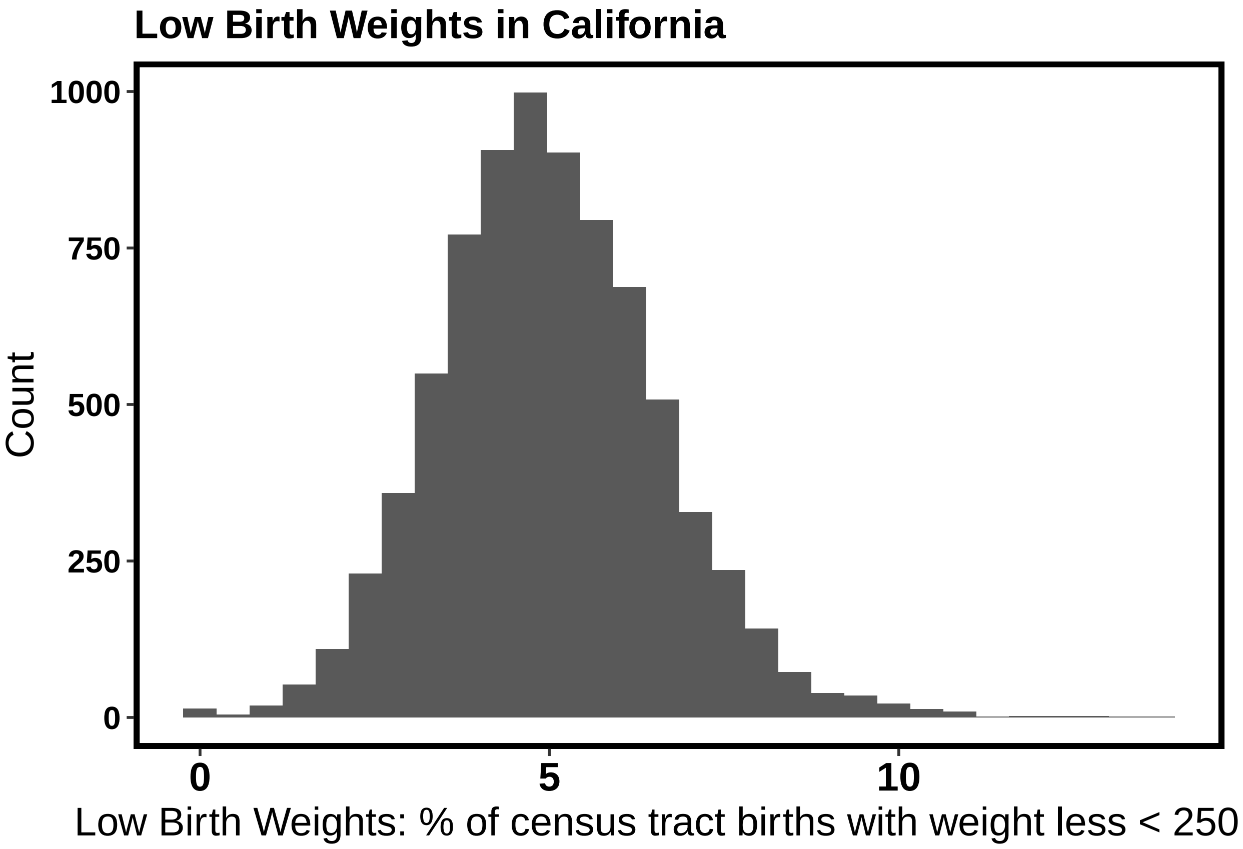
In California, Tulare county has the highest poverty with a mean of 51.4558442 percent of the county population in the census tract living below twice the federal poverty line.

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

```
#class(relevant_data$low_birth_weight)
#as.numeric(unlist(relevant_data$low_birth_weight))
# lbw_num_vec <- relevant_data %>%
#   unlist("low_birth_weight") %>%
#   as.numeric("low_birth_weight")
# lbw_num_vec
```

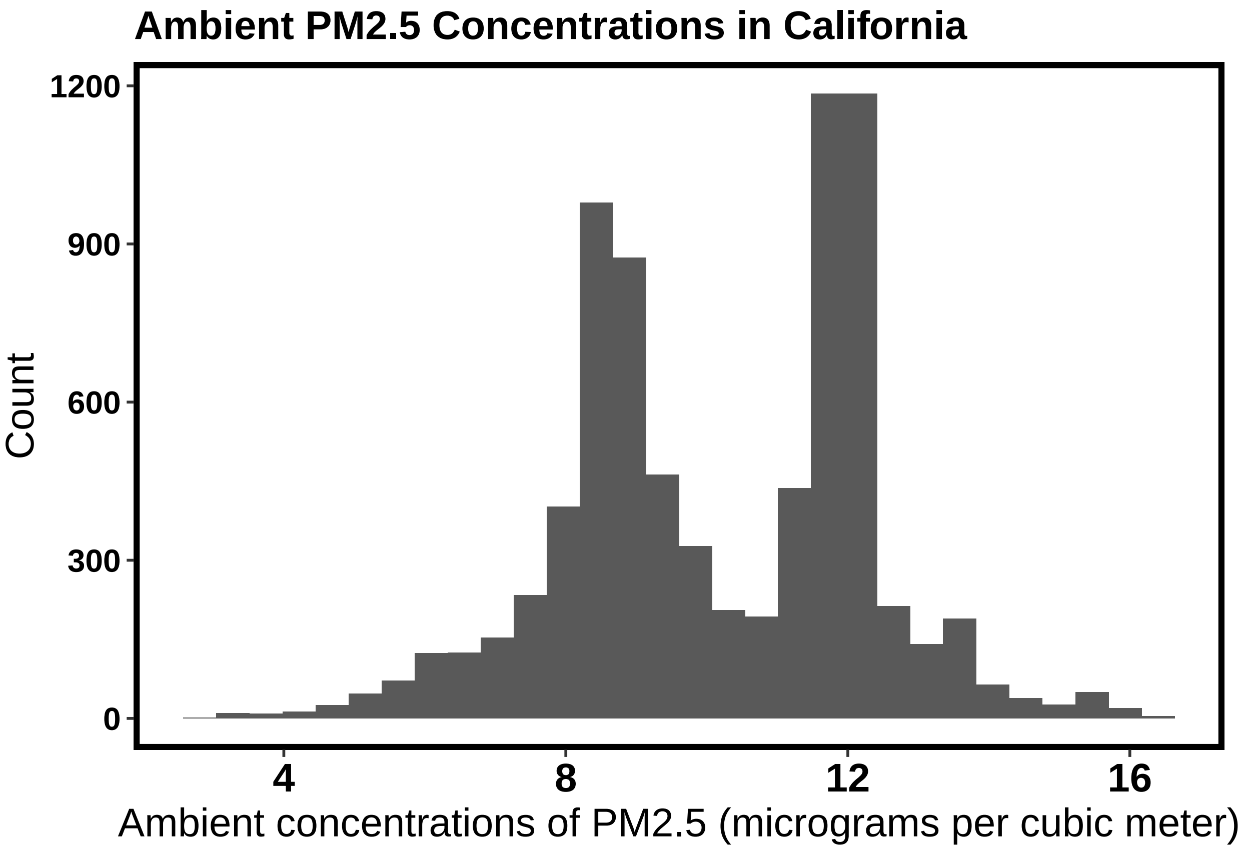
```
hist_birth_weight <- ggplot(data = relevant_data, aes(x = low_birth_weight)) +
  geom_histogram() +
  ggtitle("Low Birth Weights in California") +
   xlab("Low Birth Weights: % of census tract births with weight less < 2500g") +
   ylab("Count") +
   theme(panel.background = element_blank(),
         axis.title.x = element_text(color = "black", size = 15),
         axis.text.x = element_text(face = "bold", color = "black", size = 15),
         axis.title.y = element_text(color = "black", size = 15),
         axis.text.y = element_text(face = "bold", color = "black", size = 12),
         plot.title = element_text(color="black", size = 15, face = "bold"),
         panel.border = element_rect(colour = "black", fill = NA, size = 2))

hist_birth_weight
```



```
hist_pm2_5 <- ggplot(data = relevant_data, aes(x = pm2_5)) +
  geom_histogram() +
  ggtitle("Ambient PM2.5 Concentrations in California") +
   xlab("Ambient concentrations of PM2.5 (micrograms per cubic meter)") +
   ylab("Count") +
   theme(panel.background = element_blank(),
         axis.title.x = element_text(color = "black", size = 15),
         axis.text.x = element_text(face = "bold", color = "black", size = 15),
         axis.title.y = element_text(color = "black", size = 15),
         axis.text.y = element_text(face = "bold", color = "black", size = 12),
         plot.title = element_text(color="black", size = 15, face = "bold"),
         panel.border = element_rect(colour = "black", fill = NA, size = 2))

hist_pm2_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%?

```
pm_model <- lm_robust(formula = low_birth_weight ~ pm2_5, data = relevant_data)
pm_model
```

```
##              Estimate  Std. Error  t value     Pr(>|t|)  CI Lower  CI Upper
## (Intercept) 3.7995702 0.088577946 42.89522 0.000000e+00 3.6259337 3.9732067
## pm2_5       0.1181619 0.008401392 14.06456 2.178605e-44 0.1016929 0.1346309
##               DF
## (Intercept) 7803
## pm2_5       7803
```

```
# call coefficients, std. errors, and p-values as objects
pm_model$coefficients[1]
```

```
## (Intercept)
##     3.79957
```

```r
pm_model$coefficients[2]
```

```
##     pm2_5
## 0.1181619
```

```r
pm_model$std.error[2]
```

```
##       pm2_5
## 0.008401392
```

```r
pm_model$p.value[2]
```

```
##        pm2_5
## 2.178605e-44
```

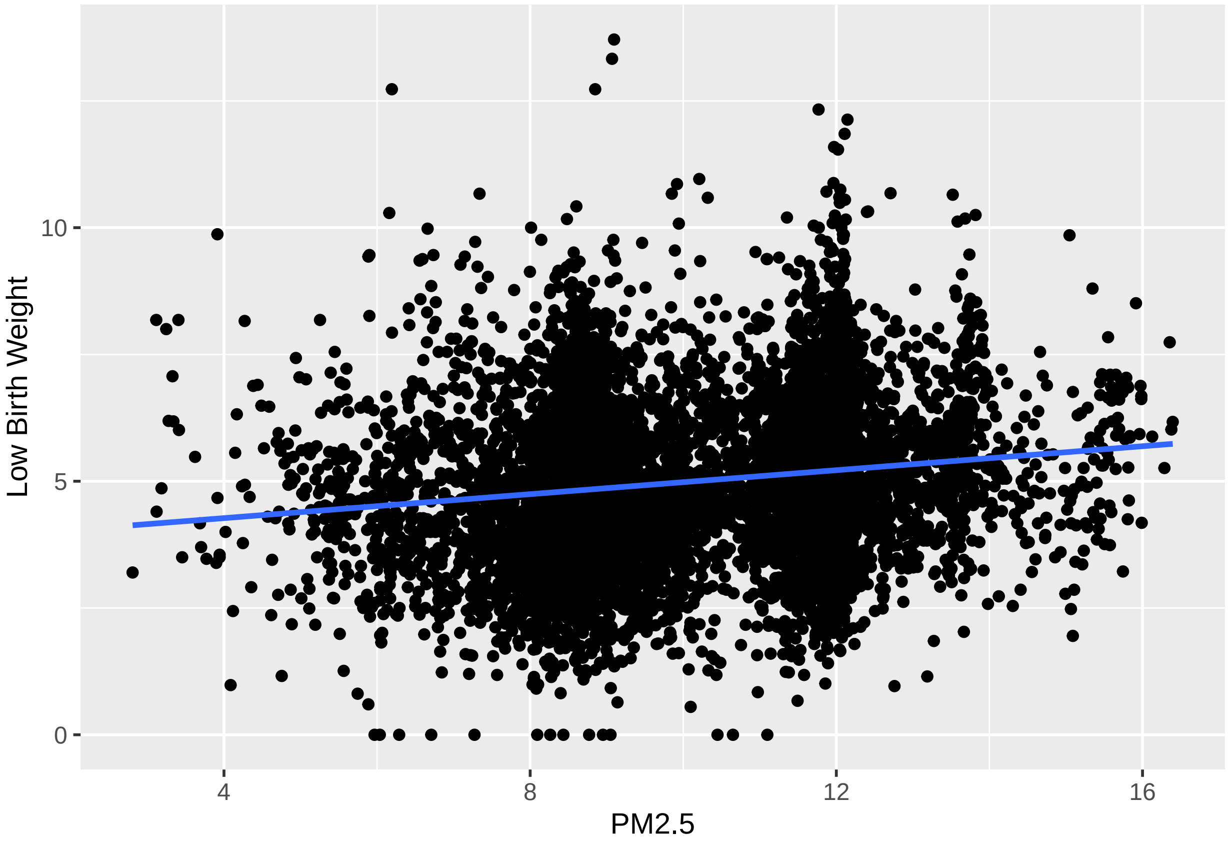
The linear equation for the relationship between PM2.5 concentration and low birth weight is: low_birth_weight = 3.7995702 + (0.1181619)number_PM_unit_increase + u

- **The estimated slope coefficient for the OLS regression of PM2.5 on Low Birth Weight is** 0.1181619
- **The heteroskedasticity-robust standard error for the slope coefficient is** 0.0084014. **We can trust that this standard error is heteroskedasticity robust because we used lm_robust() rather than just lm(), and lm_robust() uses HC2 for the standard errors as the default.**

- **The slope coefficient represents the amount of change in birth weights for each 1 unit increase in PM2.5 concentration, which is in units of micrograms per cubic meter. Since the slope coefficient is positive, the percentage of low birth rates will increase by** 0.1181619 **for every 1 microgram per cubic meter increase in PM2.5 in the ambient air.**

- **The effect of PM2_5 on Low Birth Rate is indeed statistically significant, with the PM2.5 p-value being** $2.1786047 \times 10^{-44}$, **which is much smaller than the standard threshold for significance of 0.05.**

```r
ggplot(data = relevant_data, aes(x = pm2_5, y = low_birth_weight)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  xlab("PM2.5") +
  ylab("Low Birth Weight")
```



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

```
pov_pm_model <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = relevant_data)
pov_pm_model
```

```
##                Estimate  Std. Error  t value      Pr(>|t|)   CI Lower   CI Upper
## (Intercept) 3.54374197 0.084732867 41.82252  0.000000e+00 3.37764284 3.70984111
## pm2_5       0.05910773 0.008293227  7.12723  1.115549e-12 0.04285079 0.07536468
## poverty     0.02743528 0.001002221 27.37448 1.287176e-157 0.02547066 0.02939990
##               DF
## (Intercept) 7802
## pm2_5       7802
## poverty     7802
```

```
pov_pm_model$coefficients[2]
```

```
##      pm2_5
## 0.05910773
```

```
# old pm slope coefficient - new pm slope coefficient
diff_pm_coeff <- pm_model$coefficients[2] - pov_pm_model$coefficients[2]
diff_pm_coeff
```

```
##      pm2_5
## 0.05905419
```

```
# difference is 0.05905419
```

```r
pov_pm_model$coefficients[2]
```

```
##      pm2_5
## 0.05910773
```

- **The estimated coefficient for poverty is** 0.0274353**. This means that for every 1 unit increase in poverty, which is a 1 percent increase in the population in the census tract that lives below twice the federal poverty line, the estimated low birth weight increases by** 0.0274353 **units, which is the percentage of the census tract births with weight less than 2500g, when PM2.5 is held constant.**
- **The estimated coefficient for PM2.5 is now** 0.0591077**, which is** 0.0590542 **lower than the original PM2.5 coefficient estimate of** 0.1181619**. The PM2.5 now has** 0.0590542 **much less of an impact on low birth weight with the newly added regressor poverty. These regressors are now distributing the responsiblity for the trend in low birth rate.**

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

```r
model_hypoth_test <- linearHypothesis(pov_pm_model,
                                      c("pm2_5 - poverty = 0"),
                                      white.adjust = "hc2")

p_value <- model_hypoth_test$`Pr(>Chisq)`[2]
p_value
```

```
## [1] 0.0002426369
```

Null Hypothesis: The effect of PM2.5 on Low Birth Weight = The effect of Poverty on Low Birth Weight Althernative Hypothesis: The effect of PM2.5 on Low Birth Weight =/= The effect of Poverty on Low Birth Weight

The p-value for this hypothesis test is 2.4263693×10^{-4}**, which is smaller than the standard threshold for significance of 0.05. We can indeed reject the null hypothesis that the effect of PM2.5 on Low Birth Weight is equal to the effect of Poverty on Low Birth Weight.**

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