EDS241: Assignment 1

Wylie Hampson

01/18/2022

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?

```r
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

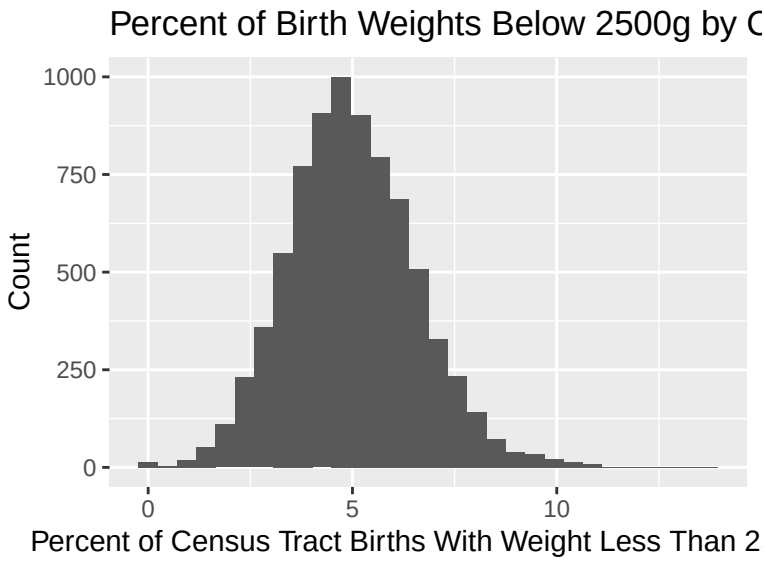
In the above summary table we can see that the California county that has the highest rate of poverty is Tulare with a poverty rate of 51.8%.

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

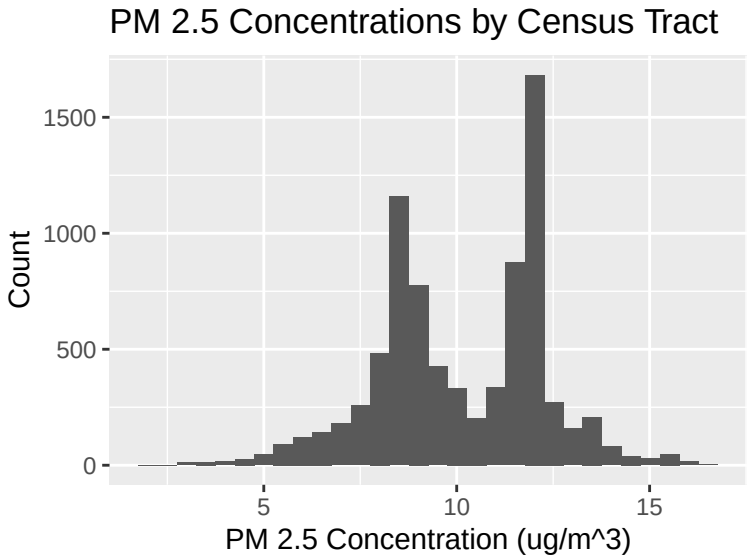
```r
birth_hist <- ggplot(data = data, aes(x = low_birth_weight)) +
  geom_histogram() +
  labs(x = "Percent of Census Tract Births With Weight Less Than 2500g",
       y = "Count",
       title = "Percent of Birth Weights Below 2500g by Census Tract")


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")

birth_hist
```




```
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, error_pos = "right")
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5, data = data)
##
## 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) 0.00000000000000000000000000000000000000000000000   3.6273   3.9746
## pm2_5       0.00000000000000000000000000000000000000000003256   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
```

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. 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 e: Suppose a new air quality policy is expected to reduce PM2.5 concentration by 2 micrograms per cubic meters. Predict the new average value of low_birth_weight and derive its 95% confidence interval. Interpret the 95% confidence interval.

```
data <- data %>%
  mutate(pm_reduced = pm2_5 - 2)

model_2 <- lm_robust(formula = low_birth_weight ~ pm_reduced, data = data)
summary(model_2, error_pos = "right")
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm_reduced, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   4.0368   0.072196   55.92
## pm_reduced    0.1179   0.008402   14.04
##                                                        Pr(>|t|) CI Lower CI Upper
## (Intercept) 0.00000000000000000000000000000000000000000000000   3.8953   4.1784
## pm_reduced  0.00000000000000000000000000000000000000000003256   0.1015   0.1344
##               DF
## (Intercept) 7806
## pm_reduced  7806
##
## Multiple R-squared:  0.02499 ,   Adjusted R-squared:  0.02486
## F-statistic:   197 on 1 and 7806 DF,  p-value: < 0.00000000000000022
```

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

```r
model_3 <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = data)
summary(model_3, error_pos = "right")
```

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

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

```r
linearHypothesis(model_3, c("pm2_5 = poverty"), white.adjust = "hc2")
```

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