EDS241: Assignment 1

Charles Hendrickson

01/20/2022

```
# Load the CalEnviroScreen 4.0 data from the California Office of Environmental Health Hazards Assessme

mydata <- read.xlsx("data/CES4.xlsx")

# Select the specific columns we will be using in our analysis.

mydata <- mydata %>%
  select("Census.Tract", "Total.Population", "California.County", "Low.Birth.Weight", "PM2.5", "Poverty

# omit all NA values from the dataset

mydata <- na.omit(mydata)
```

(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 10.19529

```
pm2.5_avg <- mydata %>%
  summarise(pm2.5_avg = mean(PM2.5))

print(pm2.5_avg)
```

```
##   pm2.5_avg
## 1  10.19529
```

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

Tulare county has the highest level of poverty in California

```
# Find the mean poverty level per county
county_poverty_means <- mydata %>%
  group_by(California.County) %>%
  summarise(mean_poverty = mean(Poverty))

# Find the highest mean poverty level out of all California counties
max(county_poverty_means$mean_poverty)
```

```
## [1] 51.45584
```

```
# Filter for the county with the max mean poverty level of 51.45584
max_poverty <- county_poverty_means %>%
  filter(mean_poverty >= 51.45584) %>%
  summarise(California.County)

print(max_poverty)
```

```
## # A tibble: 1 x 1
##   California.County
##   <chr>
## 1 "Tulare "
```

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

```r
# Histogram for PM2.5 levels
ggplot(data = mydata, aes(x= PM2.5)) +
  geom_histogram() +
  labs(title = "Distribution of Annual Mean PM2.5 Levels for California Census Tract",
       x = "PM2.5 level",
       y = "Count") +
  theme_classic()
```


Distribution of Annual Mean PM2.5 Levels
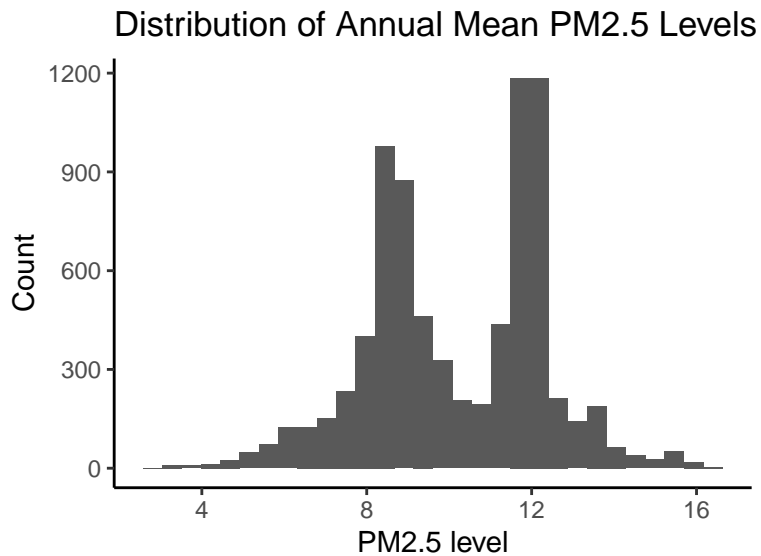
```r
# Histogram for percent low birth weight
ggplot(data = mydata, aes(x = Low.Birth.Weight)) +
  geom_histogram() +
  labs(title = "Distribution of Percent Low Birth Weight for California Census Tract",
       x = "Percent Low Birth Weight (%)",
       y = "Count") +
  theme_classic()
```

Distribution of Percent Low Birth Weight for

Percent Low Birth Weight (%)
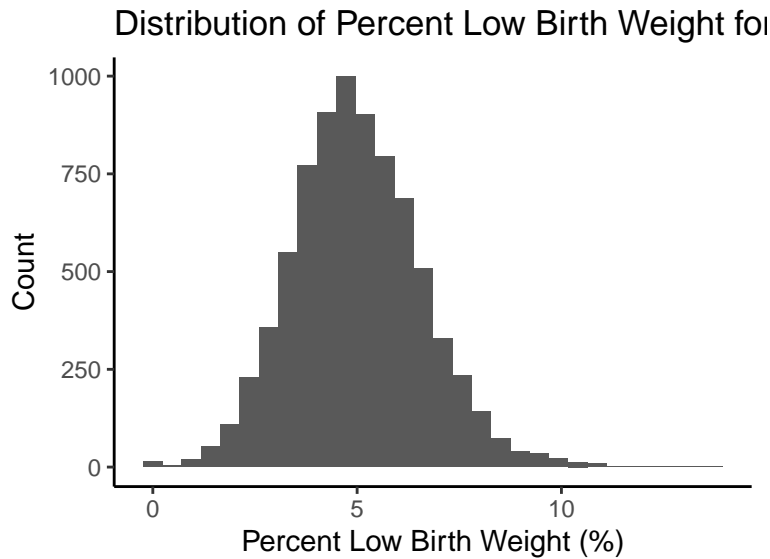
(d) Estimate a OLS regression of LowBirthWeight on PM25. Report the estimated slope coefficient and its heteroskedasticity-robust standard error. Interpret the estimated slope coefficient. Is the effect of PM25 on LowBirthWeight statistically significant at the 5%?

The estimated slope coefficient for PM2.5 is 0.118162. This means that for every one unit increase in PM2.5 levels, there is a 0.118162 percent increase in the percentage of low birth weights for the California census tract.

The standard error is 0.008334

The effect of PM2.5 on LowBirthWeight is statistically significant at the 5% level becuase our p-value is much lower than 0.05.

```
model_1 <- summary(lm_robust(Low.Birth.Weight ~ PM2.5, data = mydata))

model_1
```

```
##
## Call:
## lm_robust(formula = Low.Birth.Weight ~ PM2.5, data = mydata)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)   3.7996   0.088578   42.90 0.000e+00   3.6259   3.9732 7803
## PM2.5         0.1182   0.008401   14.06 2.179e-44   0.1017   0.1346 7803
##
## Multiple R-squared:  0.02511 ,   Adjusted R-squared:  0.02499
## F-statistic: 197.8 on 1 and 7803 DF,  p-value: < 2.2e-16
```

(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.

The estimated coefficient on Poverty is 0.02744. This means that for every one unit increase in poverty (the percent of population living below two times the federal poverty level), there is a 0.02744 percent increase in the percentage of low birth weights for the California census tract.

**The estimated coefficient on PM2.5 decreases to compared 0.1182 in part (d)

```
model_2 <- summary(lm_robust(Low.Birth.Weight ~ PM2.5 + Poverty, data = mydata))

model_2
```

```
##
## Call:
## lm_robust(formula = Low.Birth.Weight ~ PM2.5 + Poverty, data = mydata)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value   Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  3.54374   0.084733  41.823  0.000e+00  3.37764  3.70984 7802
## PM2.5        0.05911   0.008293   7.127  1.116e-12  0.04285  0.07536 7802
## Poverty      0.02744   0.001002  27.374 1.287e-157  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: < 2.2e-16
```