EDS241: Assignment 1

Charles Hendrickson

01/21/2022

```
# 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)
```

```
# Summary statistics
stargazer(mydata, type = "text", digits = 1)
```

```
##
## ====================================================================================================
## Statistic          N        Mean         St. Dev.        Min        Pctl(25)      Pctl(75)         Ma
## ----------------------------------------------------------------------------------------------------
## Census.Tract     7,805 6,054,917,124.0 26,571,997.0 6,001,400,100 6,037,262,100 6,073,016,607 6,115,
## Total.Population 7,805     4,969.9        2,219.9       507          3,539         5,954        38,
## Low.Birth.Weight 7,805       5.0           1.6          0.0           3.9           6.0          13
## PM2.5            7,805      10.2           2.1          2.8           8.6          11.9          16
## Poverty          7,805      31.3          18.1           1           16.3          44.2           93
## ----------------------------------------------------------------------------------------------------
```

(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 micrograms per cubic meter of air.

```
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?

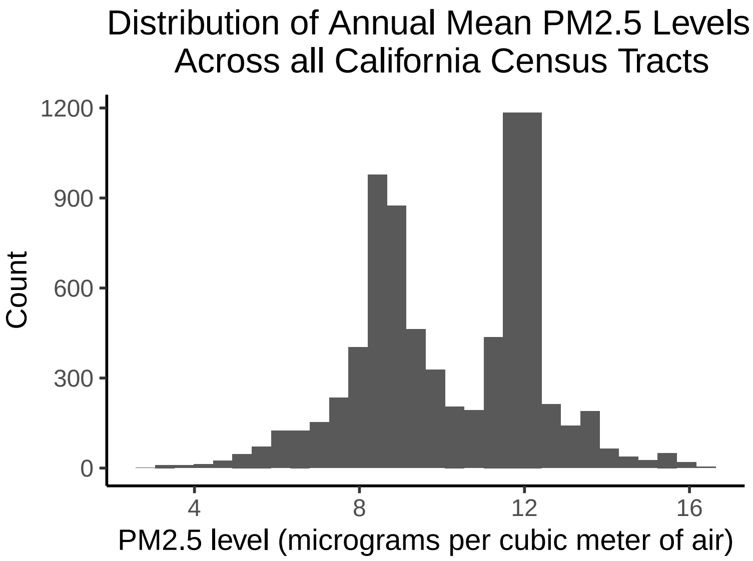
Tulare county has the highest level of poverty in California.

```
# 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_mean_poverty <- max(county_poverty_means$mean_poverty)

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

# Print the name of the county with the max mean poverty level
print(max_poverty)
```

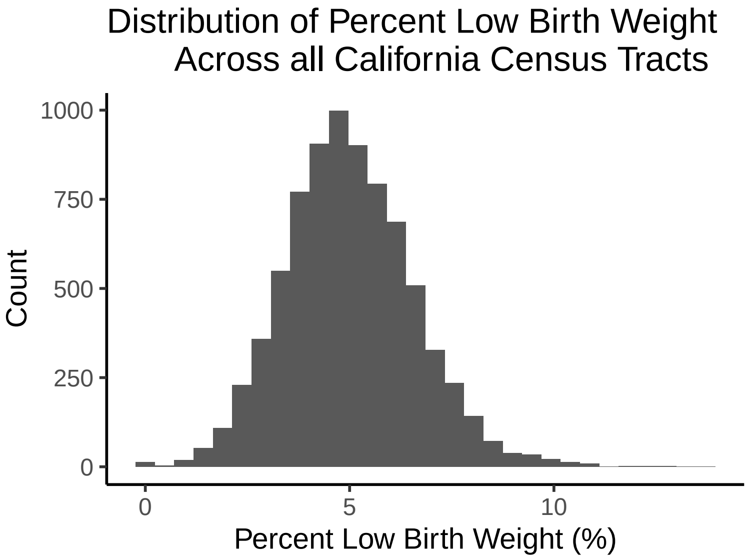
```
## # A tibble: 1 x 1
##   California.County
##   <chr>
## 1 "Tulare "
```

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

```
# Histogram for PM2.5 levels
ggplot(data = mydata, aes(x= PM2.5)) +
  geom_histogram() +
  labs(title = "Distribution of Annual Mean PM2.5 Levels
      Across all California Census Tracts",
      x = "PM2.5 level (micrograms per cubic meter of air)",
      y = "Count") +
  theme_classic()
```



```
# Histogram for percent low birth weight
ggplot(data = mydata, aes(x = Low.Birth.Weight)) +
  geom_histogram() +
  labs(title = "Distribution of Percent Low Birth Weight
      Across all California Census Tracts",
      x = "Percent Low Birth Weight (%)",
      y = "Count") +
  theme_classic()
```



(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 0.1182. This means that for the California census tract, each additional unit of concentration of PM2.5 micrograms per cubic meter of air increases the percentage of low birth weights by 0.1182 percent on average.

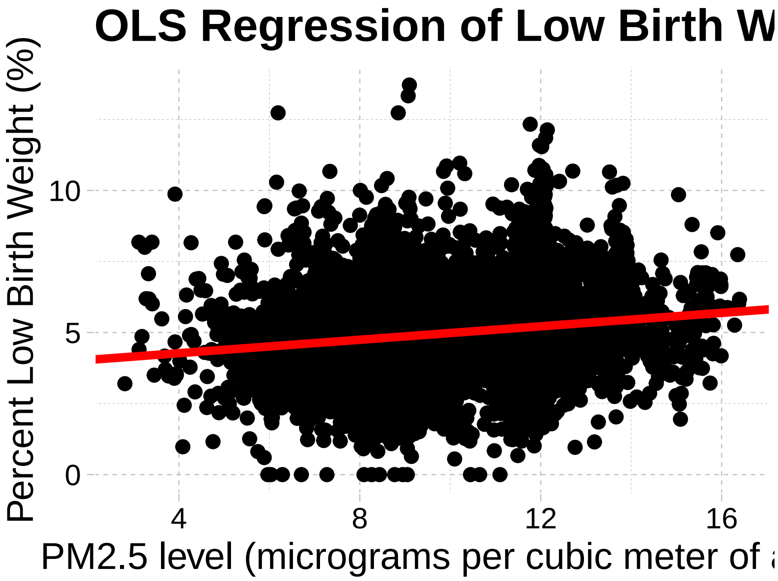
The standard error is 0.008401

The effect of PM2.5 on LowBirthWeight is statistically significant at the 5% level becuase our p-value is 2.2e-16, which is much lower than 0.05.

```
# OLS regression of Low.Birth.Weight on PM2.5
model_1 <- lm_robust(Low.Birth.Weight ~ PM2.5, data = mydata)

summary(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
```

```
# Plot of OLS Regression of Low Birth Weight on PM2.5
ggplot(mydata, aes(x = PM2.5, y = Low.Birth.Weight)) +
  geom_point(size = 2, color = "black") +
  labs(title = "OLS Regression of Low Birth Weight on PM2.5",
       x = "PM2.5 level (micrograms per cubic meter of air)",
       y = "Percent Low Birth Weight (%)") +
  ggthemes::theme_pander(base_size =
14) + geom_abline(intercept = 3.7996,
slope = 0.1182, size=1.5, color="red")
```



(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 slope coefficient on Poverty is 0.02744. This means that for the California census tract, each additional unit of poverty (the percent of population living below two times the federal poverty level) increases the percentage of low birth weights by 0.02744 on average.

In the multiple regression, the estimated slope coefficient on PM2.5 decreased from 0.1182 in part (d) to 0.05911. To explain this change, we can speculate that the true Beta2 is a positive value due to our regression showing us that poverty is positively correlated with the percentage of low birth weights. Due to omitted variables bias, when we omit the effect of poverty on the percentage of low birth weights in model_1, the regression gives that effect to PM2.5 levels, which overstates PM2.5's effect on the percentage of low birth weights and inflates the value of PM2.5's estimated slope coefficient. Thus, we observe the estimated slope coefficient for PM2.5 decrease when the poverty variable is included in model_2.

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

model_2
```

```
##                 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
```

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

We reject the null hypothesis that BetaPM2.5 = BetaPoverty because the p-value is 0.0002426,

which is much smaller than the 0.05. Also the t value for PM2.5 and Poverty is very different from eachother, which supports our rejection of the null hypothesis.

```
#
linearHypothesis(model = model_2,
                 hypothesis.matrix = 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
```