EDS 241: Assignment 1

Mia Forsline

1/21/2022

1 Setup

- load packages
- set options

1.1 Read in and clean data

- select variables of interest
    - census tract
    - California county
    - low birth: % of census tract births with weight < 2500g
    - PM2.5: ambient concentrations of PM2.5 in the census tract, in µg per m^3
    - poverty: % of population in the census tract living below twice the federal poverty line

```
data <- read_csv(here("data", "CES4.csv"))

data_clean <- data %>%
  clean_names()

data_clean <- data_clean %>%
  select(census_tract,
         california_county,
         low_birth_weight,
         pm2_5,
         poverty)
```

2 Introduction

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.

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).

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

```
mean_pm2.5 <- mean(data_clean$pm2_5) %>%
  round(digits = 2)
mean_pm2.5
```

```
## [1] 10.15
```

The mean concentration of PM2.5 across all census tracts in California is 10.15 µg/m^3.

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

- drop counties with NA values for poverty

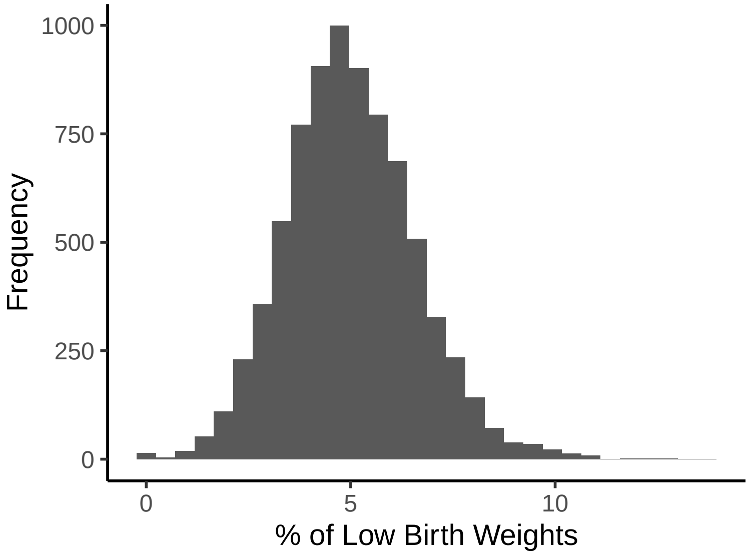
```
data_pov <- data_clean %>%
  drop_na(poverty)

max_pov_county <- subset(data_pov, poverty == max(data_pov$poverty))
max_pov_county <- max_pov_county$california_county
max_pov_county
```

```
## [1] "Ventura"
```

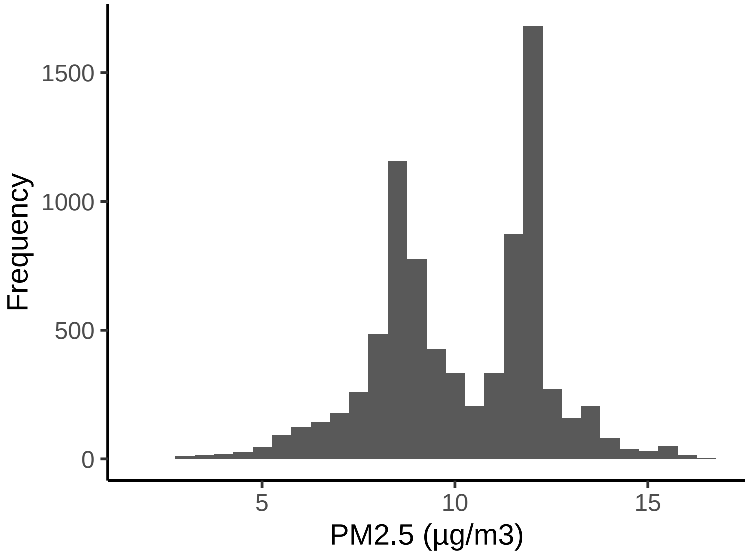
Ventura is the county with the highest level of poverty in California.

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

```
ggplot(data = data_clean) +
  geom_histogram(aes(x = low_birth_weight)) +
  theme_classic() +
  labs(y = "Frequency",
       x = "% of Low Birth Weights")
```



```r
ggplot(data = data_clean) +
  geom_histogram(aes(x = pm2_5)) +
  theme_classic() +
  labs(y = "Frequency",
       x = "PM2.5 (µg/m3)")
```



6 (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%?

```
mdl <- lm(low_birth_weight ~ pm2_5, data=data_clean)
summary(mdl)
```

```
##
## Call:
## lm(formula = low_birth_weight ~ pm2_5, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1099 -1.0535 -0.0909  0.9889  8.8363
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 3.800988   0.086842   43.77 <0.0000000000000002 ***
## pm2_5       0.117931   0.008338   14.14 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.569 on 7806 degrees of freedom
##   (227 observations deleted due to missingness)
## Multiple R-squared:  0.02499,    Adjusted R-squared:  0.02486
## F-statistic: 200.1 on 1 and 7806 DF,  p-value: < 0.00000000000000022
```

The estimated slope coefficient for PM2.5 (β_1) = 0.1179305, meaning that a 1 µg/^{3} change in PM2.5 on average associates with a 0.1179305 increase in Low Birth Rate. At the 5% significance level, the effect of PM2.5 on Low Birth Weight is statistically significant.

7 (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 LowBirthWeight and derive its 95% confidence interval. Interpret the 95% confidence interval.

```
#PM2.5 = 9.13552 + 0.21198 * LBW
x = -2
int = mdl$coefficients[1]
b1 = mdl$coefficients[2]
y = mdl$coefficients[1] + mdl$coefficients[2] * x

ci <- confint(object = mdl, parm = "pm2_5", level = 0.95)
ci
```

```
##            2.5 %    97.5 %
## pm2_5 0.1015864 0.1342746
```

Given the equation: PM2.5 = `mdl$coefficients[1]` + `mdl$coefficients[2]` * LowBirthWeight, if PM2.5 is reduced by 2, then we predict the Low Birth Rate to be 3.5651266%. The 95% confidence interval is bounded by 0.1015864 and 0.1342746, meaning we are 95% confident that the true parameter β_1 lies within this interval.

8 (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.

```
mdl <- lm(low_birth_weight ~ pm2_5 + poverty, data=data_clean)
summary(mdl)
```

```
##
## Call:
## lm(formula = low_birth_weight ~ pm2_5 + poverty, data = data_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3598 -0.9861 -0.1112  0.8975  8.4957
##
## Coefficients:
##              Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 3.5437420  0.0831090  42.640 < 0.0000000000000002 ***
## pm2_5       0.0591077  0.0081992   7.209    0.000000000000617 ***
## poverty     0.0274353  0.0009632  28.482 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.492 on 7802 degrees of freedom
##   (230 observations deleted due to missingness)
## Multiple R-squared:  0.1169, Adjusted R-squared:  0.1167
## F-statistic: 516.6 on 2 and 7802 DF,  p-value: < 0.00000000000000022
```

The estimated slope coefficient for Poverty (β_2) = 0.0274353, meaning that a 1 µg/^{3} change in PM2.5 on average associates with a 0.0274353 increase in Low Birth Rate. Compared to the prior model, the slope coefficient of PM2.5 (β_1) has decreased to 0.0591077 because Poverty helps explain some of the change in Low Birth Weight. In other words, the prior model suffered from omitted variables bias and caused us to overestimate the impact of PM2.5 alone on Low Birth Weight.

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

H0: PM2.5 = Poverty

HA: PM2.5 != Poverty

```
linearHypothesis(model = mdl, 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
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1   7803
## 2   7802  1 13.468 0.0002443 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```