EDS241: Assignment 1

Allie Cole

01/20/2022

1 Clean and plot data

The following code loads and cleans the data.

```
# Load data

library(readxl)
data1 <- read_excel("CES4.xlsx", sheet = 1, na = "NA") %>%
  clean_names() %>%
  select(census_tract,
         total_population,
         california_county,
         low_birth_weight,
         pm2_5,
         poverty) %>%
  drop_na(low_birth_weight)
```

2 Homework Questions

2.1 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

2.2 What county has the highest level of poverty in California?

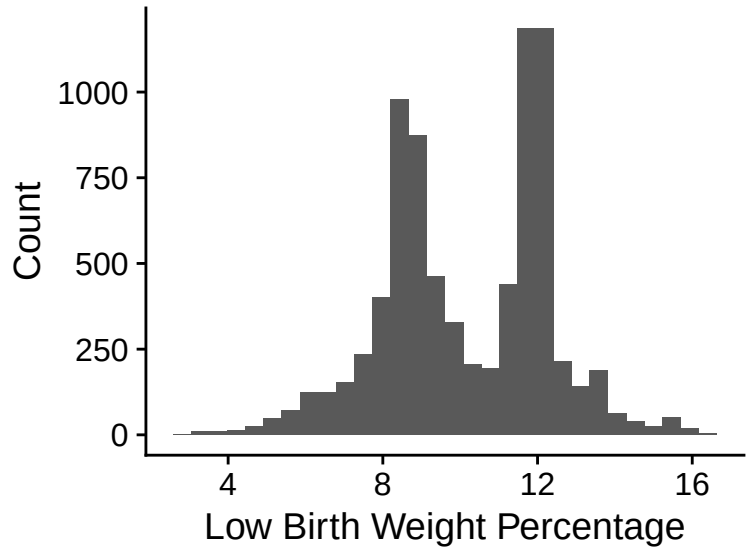
The country with the highest level of poverty in California is Ventura, and when looking at the average across all data points within the data set you can see that Tulare has the highest level of poverty.

```
average_pov <- data1 %>%
  group_by(california_county) %>%
  summarise(mean(poverty))
```

2.3 Make a histogram depicting the distribution of percent low birth weight and PM2.5.

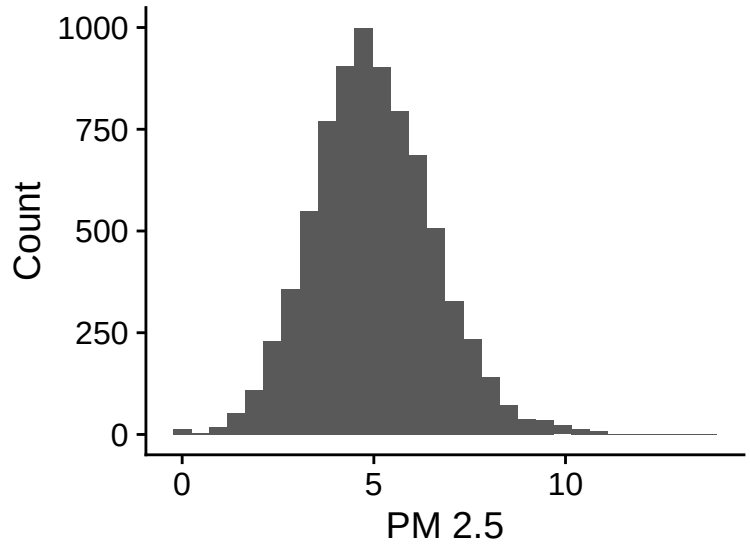
Histogram for PM2.5

```
#Histogram for PM2.5

ggplot(data = data1, aes(x = pm2_5)) +
  geom_histogram() +
  theme_cowplot(14) +
  labs(x = "Low Birth Weight Percentage",
       y = "Count")
```



Histogram for Low Birth Weight

```
#Histogram for Low Birth Weight

ggplot(data = data1, aes(x = low_birth_weight)) +
  geom_histogram() +
  theme_cowplot(14) +
  labs(x = "PM 2.5",
       y = "Count")
```



2.4 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_lbw <- lm_robust(formula = low_birth_weight ~ pm2_5, data = data1)

tidy_pm_lbw <- broom::tidy(pm_lbw)

tidy_pm_lbw %>%
  dplyr::filter(term != '(Intercept)') %>%
  dplyr::select(term, estimate, std.error, p.value) %>%
  knitr::kable() %>%
  kableExtra::kable_styling(full_width = FALSE,
                latex_options = "HOLD_position")
```

term	estimate	std.error	p.value
pm2_5	0.1179305	0.0084024	0

The estimated slope coefficient is saying the for every one unit increase in PM2.5 you will have a 0.118 increase in the percentage of low birth weights. We can say that the effect of PM25 on LowBirthWeight statistically significant at the 5% level by looking at both the p-value.

2.5 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.

This question has been pushed to the next homework assignment

2.6 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.

```
pm_lbw_p <- estimatr::lm_robust(low_birth_weight ~ pm2_5 + poverty, data = data1)
pm_lbw_p
```

```
##                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
```

```
tidy_pm_lbw_p <- broom::tidy(pm_lbw_p)

tidy_pm_lbw_p %>%
  dplyr::filter(term != '(Intercept)') %>%
  dplyr::select(term, estimate, std.error, p.value) %>%
  knitr::kable() %>%
  kableExtra::kable_styling(full_width = FALSE,
                latex_options = "HOLD_position")
```

term	estimate	std.error	p.value
pm2_5	0.0591077	0.0082932	0
poverty	0.0274353	0.0010022	0

The estimated slope coefficient is saying the for every one unit increase in poverty you will have a 0.027 increase in the percentage of low birth weights, while holding the PM2.5 levels constant. When looking at the slope coefficient of PM 2.5 we can see that it has lowered from 0.118 to 0.027. The slope is lower as now we are allowing poverty and pm 2.5 to both have an effect on low birth weights. Before only pm 2.5 was being modeled and seeming to account for the majority of the variation.

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

```
lhm <- linearHypothesis(model = pm_lbw_p, c("pm2_5=poverty"))
lhm
```

```
## 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
```

```
tidy_lhm <- broom::tidy(lhm) %>%
  dplyr::select(statistic, p.value) %>%
  dplyr::filter(statistic > 0) %>%
  dplyr::rename(chisq = statistic) %>%
  knitr::kable() %>%
  kableExtra::kable_styling(full_width = FALSE,
                latex_options = "HOLD_position")

tidy_lhm
```

chisq	p.value
13.46823	0.0002426

With this model we can reject the null hypothesis stating that the effects of poverty are the same as the effects of PM2.5 on low birth weights.