EDS241: Assignment 1

Marie Rivers

01/20/2022

In this assignment, we use air quality data in R to investigate the relationship between PM2.5 and low birth weight in California. The data came 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. Source: https://oehha.ca.gov/calenviroscreen/report/calenviroscreen-40

1 Read and Clean Data

```
data <- read_xlsx(here("data", "CES4.xlsx"), sheet = "CES4.0FINAL_results") %>%
  rename(CensusTract = "Census Tract", TotalPopulation = "Total Population", CaliforniaCounty = "Califo
  select(CensusTract, TotalPopulation, CaliforniaCounty, LowBirthWeight, PM25, Poverty) %>%
  mutate(LowBirthWeight = as.numeric(LowBirthWeight))
```

2 Question a:

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

```
mean_pm25 <- mean(data$PM25)
```

Answer: The average concentration of PM2.5 across all census tracts in California is 10.15 $\mu g/m^3$.

3 Question b:

What county has the highest level of poverty in California?

```
county_pov <- data %>%
  mutate(pov_per_capita = Poverty/TotalPopulation) %>%
  group_by(CaliforniaCounty) %>%
  summarise(mean_pov = mean(Poverty, na.rm = TRUE),
            mean_pov_per_capita = mean(pov_per_capita, na.rm = TRUE))

county_max_pov <- county_pov$CaliforniaCounty[which.max(county_pov$mean_pov)]
county_max_pov_per_capita <- county_pov$CaliforniaCounty[which.max(county_pov$mean_pov_per_capita)]
```

Answer: Tulare County has the highest level of poverty in California based on mean poverty for all census tracts in each county. I decided not to use the county with the highest level of poverty per capital (which would have been Alpine) because according to the CES4 data dictionary the `Poverty` variable represents the percent of population living below two times the federal poverty level.

Note: there are 75 census tracts with NA values for poverty

4 Question c:

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

Figure 1: Percent Low Birth Weight for California Census Tracts

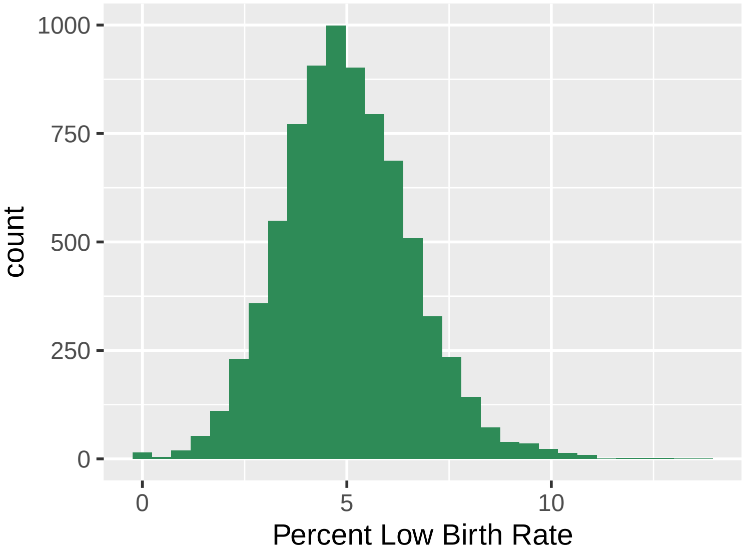


Figure 1 shows the distribution of percent low birth weights for California Census Tracts as reported by the California Office of Environmental Health Assessment (OEHHA). Data Source: CalEnviroScreen 4.0.

Figure 2: Air Quality for California Census Tracts

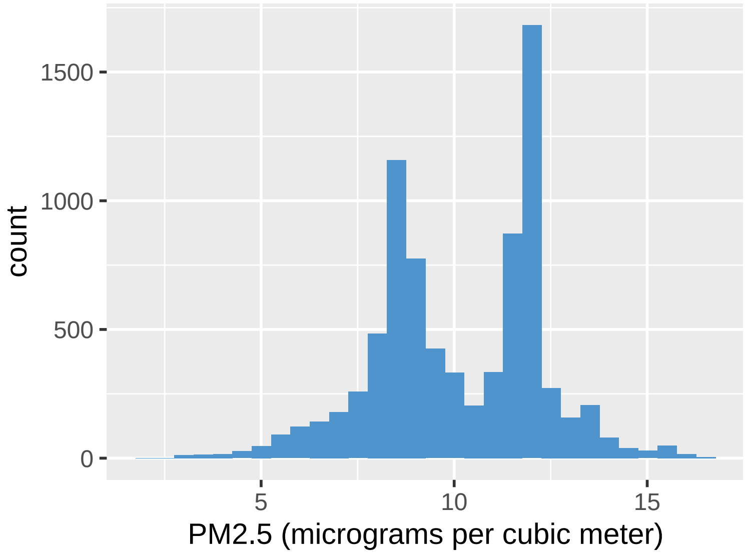


Figure 2 shows the distribution of annual mean PM 2.5 concentrations for California Census Tracts as reported by the California Office of Environmental Health Assessment (OEHHA). Data Source: CalEnviroScreen 4.0.

5 Question 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 PM2.5 on low birth weight statistically significant at the 5% level?

$$\text{percent low birth weight}_i = \beta_0 + \beta_1 \cdot PM2.5 + \varepsilon_i$$

```
ols_model <- lm_robust(formula = LowBirthWeight ~ PM25, data = data)
summary(ols_model)
```

```
##
## Call:
## lm_robust(formula = LowBirthWeight ~ PM25, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   3.8010   0.088583   42.91
## PM25          0.1179   0.008402   14.04
##                                                       Pr(>|t|) CI Lower CI Upper
## (Intercept) 0.00000000000000000000000000000000000000000000000   3.6273   3.9746
## PM25        0.00000000000000000000000000000000000000000003256   0.1015   0.1344
##               DF
## (Intercept) 7806
## PM25        7806
##
```

```
## Multiple R-squared:  0.02499 ,   Adjusted R-squared:  0.02486
## F-statistic:   197 on 1 and 7806 DF,  p-value: < 0.00000000000000022
```

```
intercept <- ols_model$coefficients[1]
intercept
```

```
## (Intercept)
##    3.800988
```

```
slope_coef <- ols_model$coefficients[2]
slope_coef
```

```
##      PM25
## 0.1179305
```

```
std_err <- ols_model$std.error[2]
std_err
```

```
##        PM25
## 0.008402393
```

```
ci_lower <- ols_model$conf.low[2]
ci_lower
```

```
##      PM25
## 0.1014596
```

```
ci_upper <- ols_model$conf.high[2]
ci_upper
```

```
##      PM25
## 0.1344015
```

```
# xxx
fit <- ols_model$fitted.values
```

Answer: The estimated slope coefficient for this model is 0.118. The estimated slope coefficient means that for a one unit increase in the concentration of PM2.5, percent low birth weight will increase by 0.008. The heteroskedasticity-robust standard error for this estimated slope coefficient is 0.008. Since the p-value is < 0.05, the effect of PM2.5 on percent low birth weight is statistically significant at the 5 % level.

```
ols_model_plot <- ggplot(data = data, aes(x = PM25, y = LowBirthWeight)) +
  geom_point() +
  geom_smooth(method = lm)
```

Figure 3: OLS Regression of Low Birth Weight on PM 2.5

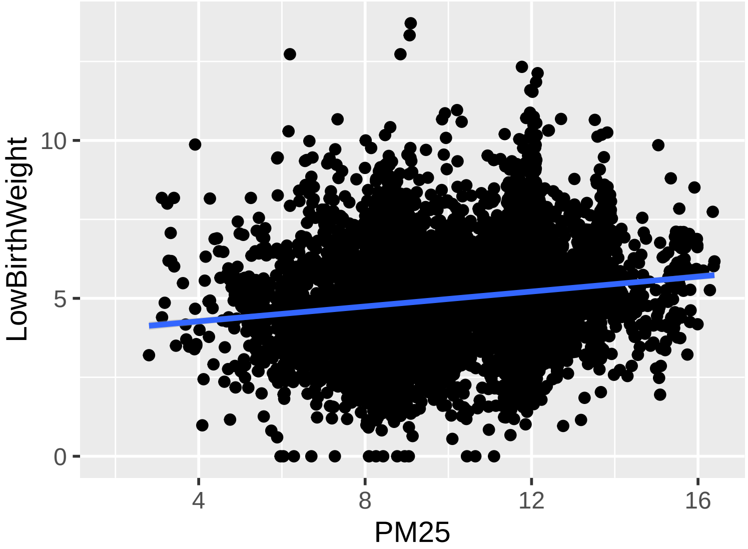


Figure 3 shows a positive correlation between low birth weight and PM 2.5 based on an Ordinary Least Squares Regression model of low birth weight on PM 2.5.

xxx…read up on "heteroskedasticity-robust standard errors"

```
# xxx...experiment chunk

# data_rm_na <- data %>%
#   na.omit(LowBirthWeight) %>%
#   mutate(model_low_birth_wt = ols_model$fitted.values)

fit <- data.frame(ols_model$fitted.values)

fit2 <- data.frame(predict(ols_model, newdata = data, se.fit = TRUE, interval = "confidence"))
```

6 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_reduced_PM <- data %>%
#   na.omit(LowBirthWeight) %>%
#   mutate(PM25_reduced = PM25 - 2) %>%
#   mutate(PM25_reduced = case_when(
#     PM25_reduced < 0 ~ 0,
#     PM25_reduced >= 0 ~ PM25_reduced)) %>%
#   mutate(test = slope_coef * PM25 + intercept) %>%
#   mutate(test_policy = slope_coef * PM25_reduced + intercept)

  #mutate(model_low_birth_wt = ols_model$fitted.values)

policy_results <- data %>%
  mutate(PM25_reduced = PM25 - 2) %>%
  mutate(PM25_reduced = case_when(
    PM25_reduced < 0 ~ 0,
    PM25_reduced >= 0 ~ PM25_reduced)) %>%
  mutate(policy_low_birth_wt = slope_coef * PM25_reduced + intercept) %>%
  mutate(policy_ci_low = (ci_lower * PM25_reduced + intercept)) %>%
  mutate(policy_ci_high = (ci_upper * PM25_reduced + intercept))
```

```
mean_policy_low_birth_wt <- mean(policy_results$policy_low_birth_wt)
mean_policy_low_birth_wt
```

```
## [1] 4.762442
```

```
mean_policy_low_birth_wt_ci_low <- mean(policy_results$policy_ci_high)
mean_policy_low_birth_wt_ci_low
```

```
## [1] 4.896725
```

```
mean_policy_low_birth_wt_ci_high <- mean(policy_results$policy_ci_high)
mean_policy_low_birth_wt_ci_high
```

```
## [1] 4.896725
```

Following an air quality policy that reduces PM2.5 concentrations by 2 microns per cubic meters, the predicted new average value of percent low birth weight is 4.762‘

The 95% confidence interval for the predicted new average value of percent low birth weight ranges from 4.897 to 4.897. This means that there is a 95% chance that this interval includes the true average value of percent low birth weight based on the effect of the new policy on PM2.5 concentrations.

xxx…state what the original low birht wt is and talk about how much it decreases due to the policy

7 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 PM2.5, compared to the regression in (d). Explain.

```
pm25_poverty_model <- lm_robust(formula = LowBirthWeight ~ PM25 + Poverty, data = data)
summary(pm25_poverty_model)
```

```
##
## Call:
## lm_robust(formula = LowBirthWeight ~ PM25 + Poverty, data = data)
##
## Standard error type:  HC2
##
## Coefficients:
##              Estimate Std. Error t value
## (Intercept)  3.54374   0.084733  41.823
## PM25         0.05911   0.008293   7.127
## Poverty      0.02744   0.001002  27.374
##
## (Intercept) 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## PM25        0.00000000001115548590940355461314964932343543315080083844215153021650621667504310607910
## Poverty     0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##             CI Lower CI Upper   DF
## (Intercept)  3.37764  3.70984 7802
## PM25         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: < 0.0000000000000022
```

```
poverty_coef <- pm25_poverty_model$coefficients[3]
poverty_coef
```

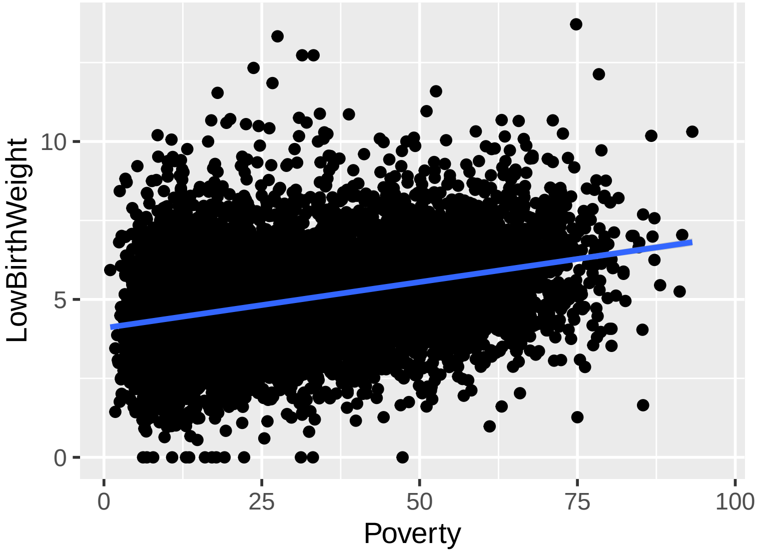
```
##    Poverty
## 0.02743528
```

```
pm25_coef_2 <- pm25_poverty_model$coefficients[2]
pm25_coef_2
```

```
##       PM25
## 0.05910773
```

The estimated coefficient on poverty is 0.027. This mean that for a one unit increase in poverty, percent low birth weighth will increase by 0.027 while holding PM2.5 constant xxx…check/clean up this interpretation

In this model, the estimated coefficient on PM2.5 is lower than the coefficient estimated with the model that did not include poverty. This suggests that PM2.5 is not the only variable to influences low birth weight.



8 Question g:

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

null hypothesis: There is no difference in the estimated slope coefficients for PM2.5 and poverty.

$$H_0 : \beta_{1,PM2.5} - \beta_{2,poverty} = 0$$

alternative hypothesis: There is a difference in the estimated slope coefficient for PM2.5 and poverty.

$$H_A : \beta_{1,PM2.5} - \beta_{2,poverty} \neq 0$$

```
se_pm25 <- pm25_poverty_model$std.error[2]
se_pm25
```

```
##        PM25
## 0.008293227
```

```
se_poverty <- pm25_poverty_model$std.error[3]
se_poverty
```

```
##     Poverty
## 0.001002221
```

```
n_pm25 <- length(data$LowBirthWeight[!is.na(data$LowBirthWeight)])
n_pm25
```

```
## [1] 7808
```

```
n_poverty <- data %>%
  filter(!is.na(Poverty)) %>%
  filter(!is.na(LowBirthWeight)) %>%
  nrow()
n_poverty
```

```
## [1] 7805
```

```
# xxx
# data_pov <- data %>%
#   filter(!is.na(Poverty)) %>%
#   filter(!is.na(LowBirthWeight)) %>%
#   nrow()
# data_pov
```

```
# not sure if this is the correct method
se <- sqrt(((se_pm25^2) / n_pm25) + ((se_poverty^2) / n_poverty))
se
```

```
##          PM25
## 0.00009453729
```

```
point_est <- pm25_coef_2 - poverty_coef
point_est
```

```
##       PM25
## 0.03167245
```

```
z_score <- (point_est - 0) / se
z_score
```

```
##    PM25
## 335.026
```

```
p_value <- 2 * pnorm(point_est, mean = 0, sd = se, lower.tail = FALSE)
p_value
```

```
## PM25
##    0
```

Since the p-values is <0.05 we reject the null hypothesis that the effect of PM2.5 is equal to the effect of poverty. xxx…check this

```
# xxx
colSums(!is.na(data))
```

```
##      CensusTract  TotalPopulation CaliforniaCounty   LowBirthWeight
##             8035             8035             8035             7808
##             PM25          Poverty
##             8035             7960
```

xxx…delete everything below here

9 Clean and plot data

The following code loads and cleans the data.

```
# Load data

data("mtcars")
raw_data <- mtcars

# Clean data

## Add model names as a column
## [this is just an example manipulation, I rarely assign rownames to a column]

clean_data <- tibble::rownames_to_column(raw_data, "model")
```

The code chunk below shows how to produce a scatter plot of MPG against weight.

```r
# Plot 1

plot_1 <- ggplot(clean_data, aes(y=mpg, x = wt))+
  geom_point()+
  theme_cowplot(14)+
  labs(x = "Weight (1000 lbs)", y = "Miles per gallon")
```

Figure 1: MPG and vehicle weight

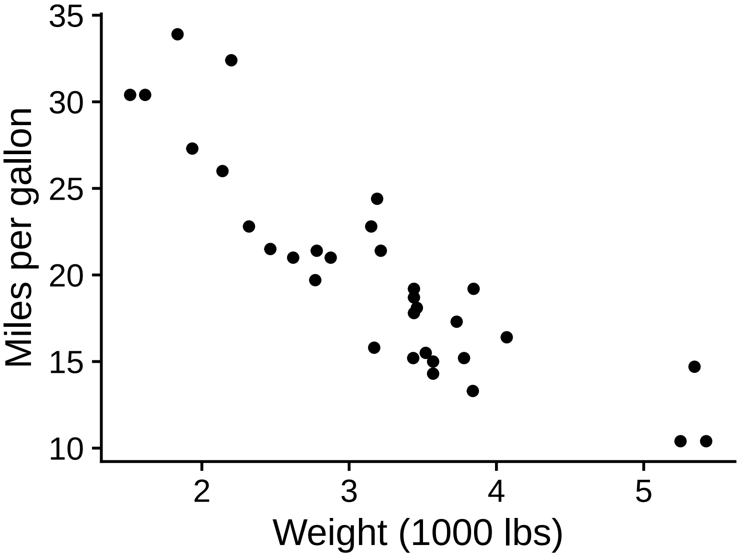


Figure 1 shows the expected negative relationship between vehicle weight and MPG.

10 Run and interpret regression models

In order to more formally analyze the relationship between MPG, vehicle weight, and cylinders we estimate the following regression:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i \tag{1}$$

where Y_i is MPG for vehicle model i, X_{1i} is the vehicle weight, X_{2i} is the number of cylinders in the engine, and u_i the regression error term. We will consider a regression including only vehicle weight, and a regression including vehicle weight and number of cylinders.

In R, we run the following code:

```
model_1 <- lm(mpg ~ wt, data=clean_data)
model_2 <- lm(mpg ~ wt + cyl, data=clean_data)
```

Table 1 shows the estimated coeffients from estimating equation (1).

Table 1: MPG and vehicle weight

	MPG	
	(1)	(2)
Weight (1000 lbs)	-5.344^{***}	-3.191^{***}
	(0.559)	(0.757)
Cylinders		-1.508^{***}
		(0.415)
Observations	32	32
R^2	0.753	0.830
Note:	*p<0.1; **p<0.05; ***p<0.01	

In model (1), the estimated β_1 coefficient implies that a 1000 pound increase in vehicle weight reduces miles per gallon by 5.3 miles. Adding the number of cylinders in model (2) reduces $\hat{\beta}_1$ from -5.3 to -3.2.