Assignment 1 EDS241

Guillermo Romero

2023-02-27

University of California, Santa Barbara Olivier Deschenes Bren School of Env. Science & Management Winter 2023 EDS 241 Assignment 1 Due on 2/27/23

Turn in your Markdown pdf on Gauchospace in the "Assignment Turn in Area"

The data for this assignment are taken 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 The full data are contained in the file CES4.xls, which is available on Gauchospace (note that the Excel file has three "tabs" or "sheets"). The data is in the tab "CES4.0FINAL_results" and "Data Dictionary" contains the definition of the variables.

For the assignment, you will need the following variables: CensusTract, TotalPopulation, 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), Poverty (percent of population in the census tract living below twice the federal poverty line), and LinguisticIsolation (percent of households in the census tract with limited English speaking).

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0     v purrr   0.3.4
## v tibble  3.1.8     v dplyr   1.0.9
## v tidyr   1.2.0     v stringr 1.4.0
## v readr   2.1.2     v forcats 0.5.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
library(estimatr)
library(stargazer)
```

```
##
## Please cite as:
##
##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(ggplot2)
library(flextable)
```

```
##
## Attaching package: 'flextable'
##
## The following object is masked from 'package:purrr':
##
##     compose
```

```
ces4_dat <- read_excel("CES4.xlsx",
                       na = "NA") %>%
  janitor::clean_names() %>%
  select(c(census_tract,
           total_population,
           low_birth_weight,
           pm2_5,
           poverty,
           linguistic_isolation))
```

```
avg_pm_all <- mean(ces4_dat$pm2_5)
avg_pm_all
```

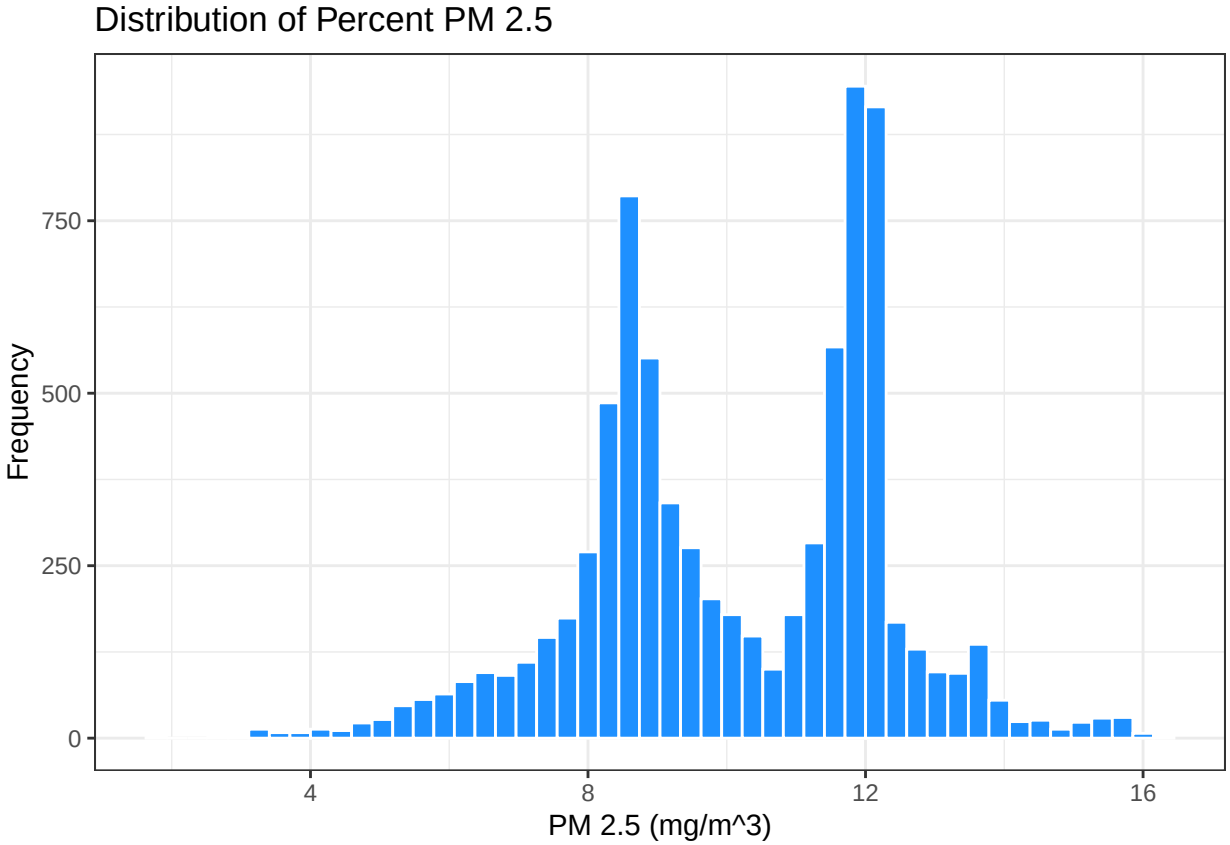
(a) What is the average concentration of PM2.5 across all census tracts in California?

```
## [1] 10.1527
```

> **The average concentration of PM2.5 across all census tracks in California is 10.15 micrograms per cubic meter.**

```
ggplot(ces4_dat, aes(x = pm2_5)) +
  geom_histogram(bins = 50, color = "white", fill = "dodgerblue") +
  labs(title = "Distribution of Percent PM 2.5", x = "PM 2.5 (mg/m^3)", y = "Frequency") +
  theme_bw()
```

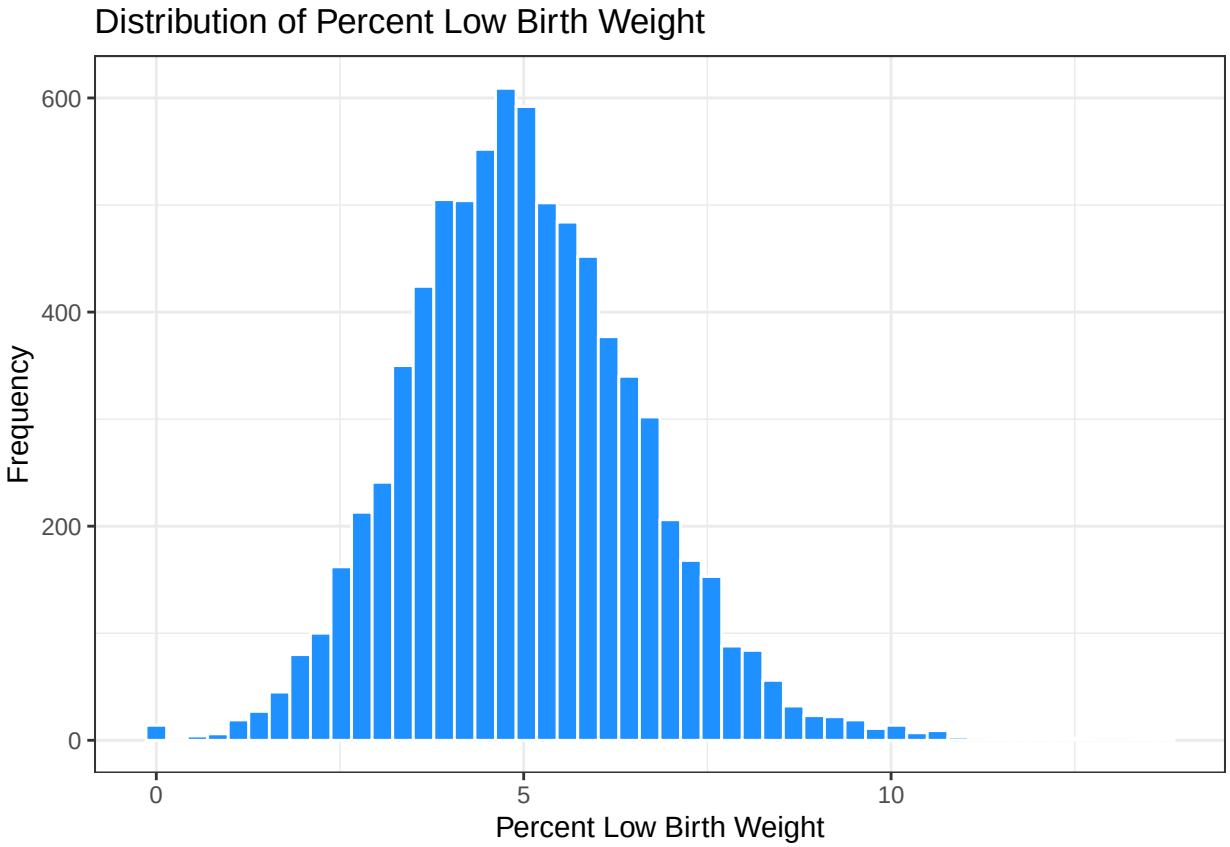
(b) Make a histogram depicting the distribution of percent low birth weight and PM2.5.




```
low_birth <- ces4_dat %>% mutate(low_b = as.numeric(ces4_dat$low_birth_weight))

ggplot(low_birth, aes(x = low_b)) +
  geom_histogram(bins = 50, color = "white", fill = "dodgerblue") +
  labs(title = "Distribution of Percent Low Birth Weight", x = "Percent Low Birth Weight", y = "Frequen
  theme_bw()
```

```
## Warning: Removed 227 rows containing non-finite values (‘stat_bin()‘).
```




```
mod <- lm_robust(formula = low_birth_weight ~ pm2_5, data = ces4_dat)

summary(mod)
```

(c) Estimate an 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% level?

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5, data = ces4_dat)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)   3.8010   0.088583   42.91 0.000e+00   3.6273   3.9746 7806
## pm2_5         0.1179   0.008402   14.04 3.256e-44   0.1015   0.1344 7806
##
## Multiple R-squared:  0.02499 ,   Adjusted R-squared:  0.02486
## F-statistic:   197 on 1 and 7806 DF,  p-value: < 2.2e-16
```

The estimated slope coefficient is 01.2 . This means that for every 1 mg/cm^3 increase in NOX concentration, the percent of census tract births with weight less than 2500g will increase by 0.008. The effect of PM2.5 on low birth weight percentage is statistically significant since the p-value is lower than the signficant level of 5%.

```
pm_reduced <- mean(ces4_dat$pm2_5) - 2
pm_red_df <- data.frame(pm2_5 = pm_reduced)
pm_reduced
```

(d) 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.

Thescript"LinearPrediction.R"availableonGauchospacewillbehelpfulforthis.

```
## [1] 8.1527
```

```
pm_red_df
```

```
##    pm2_5
## 1 8.1527
```

```
prediction <- predict(mod, newdata = pm_red_df, se.fit = TRUE, interval = 'confidence')
prediction
```

```
## $fit
##          fit      lwr      upr
## [1,] 4.76244 4.712522 4.812358
##
## $se.fit
##          1
## 0.02546467
```

```
# CI 4.71 - 4.81

# THE AVERAGE LOW BIRTH WEIGHT % IS 4.76 -------------------
avg <- mean(prediction$fit, na.rm = TRUE)
avg
```

```
## [1] 4.76244
```

The new average value of low birth weight is 4.76 percent. The 95% confidence interval is 4.71 and 4.81.

```
ols_bv_r <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty ,
                      data = ces4_dat,
                      se_type = "HC2",
                      alpha = 0.05)

summary(ols_bv_r)
```

(e) 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.

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = ces4_dat,
##     se_type = "HC2", alpha = 0.05)
##
## 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
```

```
est <- round(ols_bv_r$coefficients[2:3],2)
est
```

```
##   pm2_5 poverty
##    0.06    0.03
```

```
coef <- round(ols_bv_r$coefficients[3],2)
coef
```

```
## poverty
##    0.03
```

The estimate slope coefficient of property is “, coef,”. If PM 2.5 is 0 the effect of poverty on low birth weight percentage is 0.03 percent of population in the census tract population living twice below the federal poverty line per on percentage increase in low birth weight. The coefficien of PM 2.5 decreased from 0.12 to 0.06 it can be expected as adding another regressor to the analysis now has the impacts of 2 variables now. The change could be an impact of the ommited variable bais.

```
i_var <- ifelse(ces4_dat$linguistic_isolation < 6.9, 0, 1)
ces4_dat_new <- cbind(ces4_dat, i_var)
```

```r
model_mr <- lm_robust(formula = low_birth_weight ~ pm2_5 + poverty + i_var,
                      data = ces4_dat_new,
                      se_type = "HC2",
                      alpha = 0.05)

summary(model_mr)
```

(f) Create an indicator variable equal to 1 if the census tract is above the median LinguisticIsolation (6.9), and equal to 0 otherwise. Add this indicator variable to regression model used in (e) and interpret the estimated coefficient on the indicator variable.

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5 + poverty + i_var,
##     data = ces4_dat_new, se_type = "HC2", alpha = 0.05)
##
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)   3.6285   0.085344  42.516 0.000e+00  3.46117  3.79577 7599
## pm2_5         0.0467   0.008440   5.534 3.242e-08  0.03016  0.06325 7599
## poverty       0.0240   0.001183  20.278 4.430e-89  0.02168  0.02632 7599
## i_var         0.2900   0.041324   7.018 2.445e-12  0.20900  0.37101 7599
##
## Multiple R-squared:  0.1259 ,    Adjusted R-squared:  0.1255
## F-statistic: 360.4 on 3 and 7599 DF,  p-value: < 2.2e-16
```

The estimated coefficient for the indicator variable i_var is 0.29. The p-value for the coefficient is 2.445e-12, which is less than the significance level of 0.05, indicating that the coefficient is statistically significant. The estimated coefficient of 0.29 for the indicator variable i_var means that, on average, census tracts above the median linguistic isolation (0.69) have 0.29 higher low birth weight rate than census tracts below median linguistic isolation, after holding at constant PM2.5 and poverty. This indicates that the linguistic isolation could be a risk factor for low birth weight, independent of air pollution(pm2.5) and poverty variables.