EDS241: Assignment 1

Lewis 1

2023-02-27

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).

Read in and clean the data The following code loads and cleans the data.

```
# read in data, clean the column names, and select
# variables of interest
env_health <- read_csv(here("Assignment_1/CES4.csv")) %>%
    clean_names() %>%
    select(census_tract, total_population, low_birth_weight,
        pm2_5, poverty, linguistic_isolation)
```

```
# make sure that each row is a unique census tract
length(unique(env_health$census_tract))
```

(a) What is the average concentration of PM2.5 across all census tracts in California? [1] 8035

```
# paste a statement that includes the mean PM2.5 across all
# census tracts in California.
print(paste("The average concentration of PM2.5 across all census tracts in California is",
    round(mean(env_health$pm2_5), 3)))
```

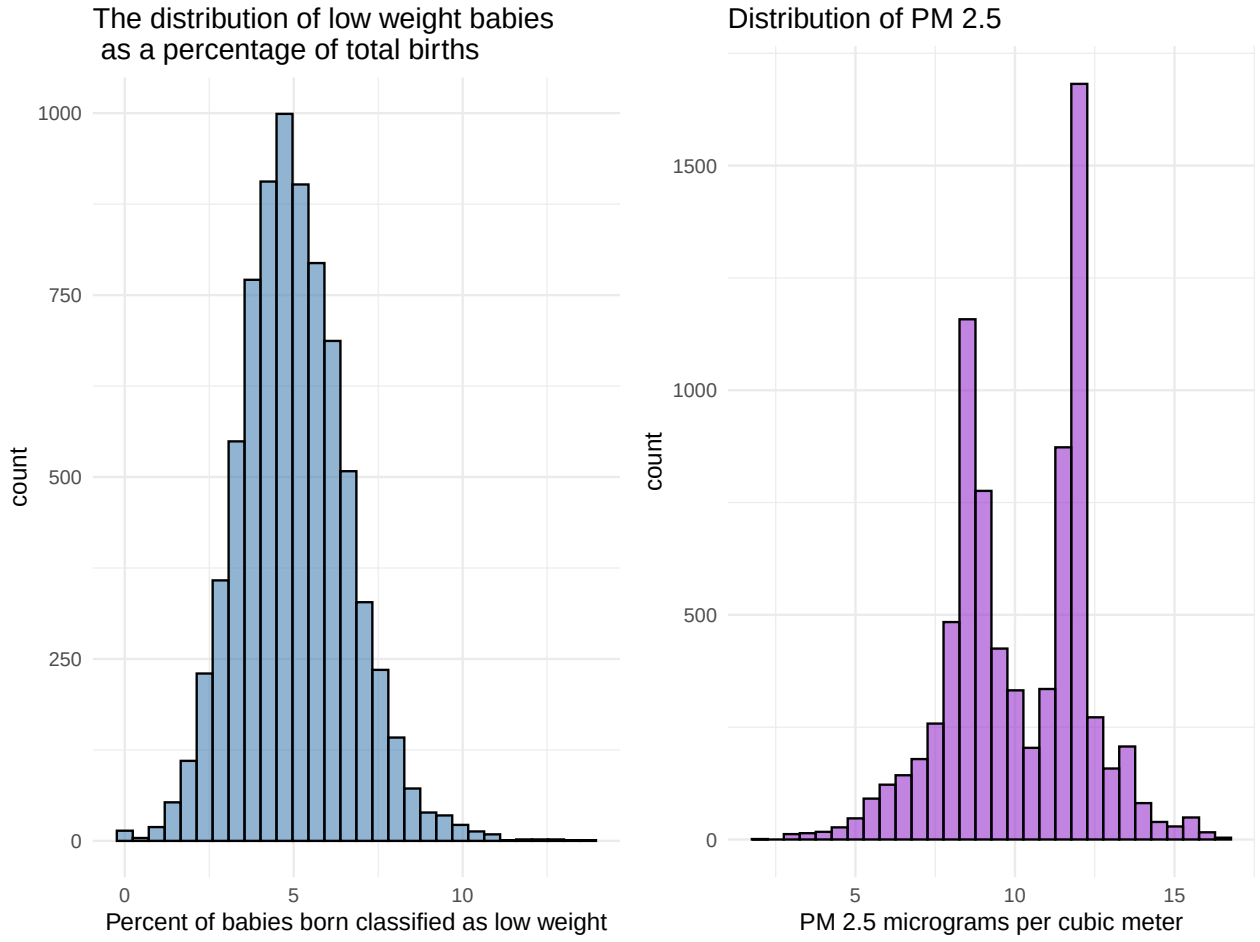
[1] "The average concentration of PM2.5 across all census tracts in California is 10.153"

```
# convert low birth weight to be numeric
env_health$low_birth_weight <- as.numeric(env_health$low_birth_weight)

# plot distribution of low birth weight
birth_wt_hist <- ggplot(data = env_health, aes(x = low_birth_weight)) +
    geom_histogram(stat = "bin", color = "black", fill = "steelblue",
```

```
        alpha = 0.6) + theme_minimal() + labs(x = "Percent of babies born classified as low weight",
    title = "The distribution of low weight babies \n as a percentage of total births")

# plot distribution of pm2.5
pm_hist <- ggplot(data = env_health, aes(x = pm2_5)) + geom_histogram(stat = "bin",
    color = "black", fill = "darkorchid", alpha = 0.6) + theme_minimal() +
    labs(x = "PM 2.5 micrograms per cubic meter", title = "Distribution of PM 2.5")
```

(b) Make a histogram depicting the distribution of percent low birth weight and PM2.5.
Figure 1: Low birth weight and PM 2.5 Distributions

```
# plot the histograms in one chart
gridExtra::grid.arrange(birth_wt_hist, pm_hist, ncol = 2)
```



(c) Estimate an OLS regression of LowBirthWeight on PM2.5. 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?
[1] "The estimated slope coefficient for PM 2.5 is 0.11793 and the corresponding standard error for this value is 0.0084"

```
# summary table for the model
summary(birth_weight_pm25_mod)
```

```
##
```

```
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5, data = env_health,
##     se_type = "HC1", alpha = 0.05)
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)   3.8010   0.088573   42.91
## pm2_5         0.1179   0.008401   14.04
##                                                          Pr(>|t|) CI Lower CI Upper
## (Intercept) 0.00000000000000000000000000000000000000000000000000   3.6274   3.9746
## pm2_5       0.00000000000000000000000000000000000000000003187   0.1015   0.1344
##               DF
## (Intercept) 7806
## pm2_5       7806
##
## Multiple R-squared:  0.02499 ,   Adjusted R-squared:  0.02486
## F-statistic:   197 on 1 and 7806 DF,  p-value: < 0.00000000000000022
```

The estimated slope coefficient for PM 2.5 is 0.11793 and the corresponding standard error for this value is 0.0084. For each 1 microgram per cubic meter increase in the PM 2.5 value, our model predicts that the percentage of babies qualified as low-weight will increase by 0.12%. With a p value of less than 2.2e-16, this effect is statistically significant at the 5% significance level.

```
# finding the mean pm2.5 value if it was 2 micrograms per
# cubic meter lower than before
pm2_5_lower <- mean(env_health$pm2_5 - 2)

# use predict() to find the average value and confidence
# interval values given this new mean PM2.5
predict(birth_weight_pm25_mod, newdata = list(pm2_5 = pm2_5_lower),
    se.fit = TRUE, interval = "confidence")
```

(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. [The script "LinearPrediction.R" available on Gauchospace will be helpful for this.]

```
## $fit
##          fit      lwr      upr
## [1,] 4.76244 4.712526 4.812354
##
## $se.fit
##          1
## 0.02546284
```

```
# RESULTS fit lwr upr [1, ] 4.76244 4.712526 4.812354
```

Our model predicts that the new average percentage of babies born with a low birth weight would be 4.76%.

I am 95% confident that the 2 microgram per cubic meter decreases in PM2.5 will result in a new average percentage of babies born with a low birth weight between 4.712526% and

4.812354%.

```r
# adding poverty variable to the model
birth_wt_pm25_poverty <- lm_robust(low_birth_weight ~ pm2_5 +
    poverty, data = env_health, se_type = "HC1", alpha = 0.05)

# summarizing the new model
summary(birth_wt_pm25_poverty)
```

(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 = env_health,
##     se_type = "HC1", alpha = 0.05)
##
## Standard error type:  HC1
##
## Coefficients:
##             Estimate Std. Error t value
## (Intercept)  3.54374   0.084724  41.827
## pm2_5        0.05911   0.008292   7.128
## poverty      0.02744   0.001002  27.378
##
## (Intercept) 0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## pm2_5       0.0000000000110767488020121299121283477957700423300116399083492524368921294808387756347
## poverty     0.00000000000000000000000000000000000000000000000000000000000000000000000000000000000000
##             CI Lower CI Upper   DF
## (Intercept)  3.37766  3.70982 7802
## pm2_5        0.04285  0.07536 7802
## poverty      0.02547  0.02940 7802
##
## Multiple R-squared:  0.1169 ,    Adjusted R-squared:  0.1167
## F-statistic: 494.9 on 2 and 7802 DF,  p-value: < 0.00000000000000022
```

Holding PM 2.5 levels constant, our model predicts that every one percent increase in the population living below two times the federal poverty level will be associated with a 0.02744 percent increase in the percentage of babies born with a low birth weight.

The estimated coefficient for PM 2.5 decreases in this new model from 0.1179 when just PM 2.5 was in the model to 0.05911 here. It appears as if poverty was causing omitted variable bias in the first model. Poverty is positively associated with low birth weights. Based on how the coefficient for PM2.5 changed, poverty must be positively associated with PM2.5 as well.

```r
# adding an indicator variable for lingusitic isolation
env_full <- env_health %>%
    mutate(linguistic_isolation = as.numeric(linguistic_isolation)) %>%
    mutate(ling_iso_over_6point9 = ifelse(linguistic_isolation >
        6.9, 1, 0))
```

```r
# full model
birth_wt_full_model <- lm_robust(low_birth_weight ~ pm2_5 + poverty +
    ling_iso_over_6point9, data = env_full, se_type = "HC1",
    alpha = 0.05)

# summarizing the full model
summary(birth_wt_full_model)
```

(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 + ling_iso_over_6point9,
##     data = env_full, se_type = "HC1", alpha = 0.05)
##
## Standard error type:  HC1
##
## Coefficients:
##                       Estimate Std. Error t value
## (Intercept)            3.63039   0.085357  42.532
## pm2_5                  0.04675   0.008439   5.539
## poverty                0.02391   0.001186  20.159
## ling_iso_over_6point9  0.29387   0.041394   7.099
##
## (Intercept)           0.0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## pm2_5                 0.0000003140117084361698811528158304696989322479794282116927206516265869140625
## poverty               0.0000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
## ling_iso_over_6point9 0.0000000000136723135937395562242668301076128266552736634231735024513909593224
##                       CI Lower CI Upper   DF
## (Intercept)            3.46307  3.79771 7599
## pm2_5                  0.03020  0.06329 7599
## poverty                0.02158  0.02623 7599
## ling_iso_over_6point9  0.21273  0.37501 7599
##
## Multiple R-squared:  0.126 , Adjusted R-squared:  0.1257
## F-statistic: 360.9 on 3 and 7599 DF,  p-value: < 0.00000000000000022
```

When the percent linguistic isolation is over 6.9, our model predicts that the percentage of babies born with a low birth weight will be 0.29387 percent higher.