EDS241: Assignment template/example

Vincent Thivierge

01/17/2022

1 Load data

Reading in the data and selecting the variables of interest

```
df <- read_csv(here("data", "CES4.csv")) %>%
  clean_names() %>%
  select(census_tract, total_population, california_county, low_birth_weight, pm2_5, poverty) %>%
  rename(pm25 = pm2_5)
```

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

```
avg_pm25 <- mean(df$pm25)
```

The average concentration of PM2.5 across all census tracts in California is 10.1526999 $\mu g/m^3$

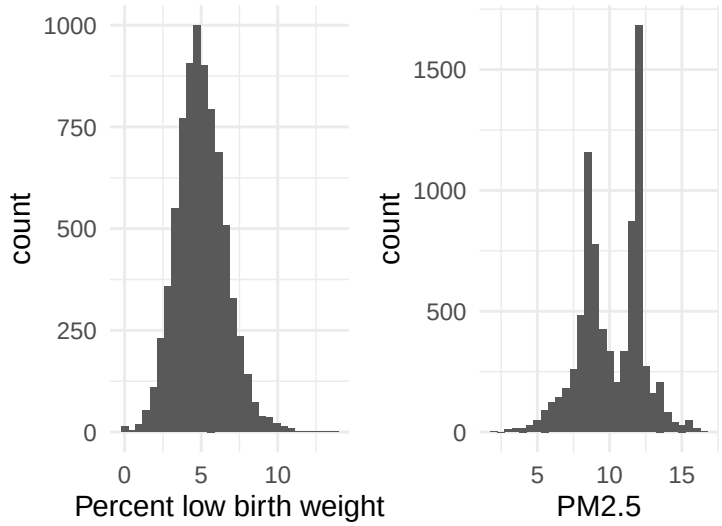
3 (b) What county has the highest level of poverty in California?

```
highest_poverty <- df %>%
  filter(poverty == max(df$poverty, na.rm = TRUE))

highest_poverty_county <- highest_poverty$california_county
```

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

```
lbw_plot <- ggplot(df, aes(x = low_birth_weight)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "Percent low birth weight")
```

```
pm25_plot <- ggplot(df, aes(x = pm25)) +
  geom_histogram() +
  theme_minimal() +
  labs(x = "PM2.5")

lbw_plot + pm25_plot
```



5 (d) OLS regression of `low_birth_weight` on `pm25`

```
model <- lm_robust(low_birth_weight ~ pm25, df)

model %>% summary()
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm25, data = df)
##
## 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
```

```
beta1_hat <- model$coefficients[["pm25"]]
```

The estimated slope coefficient is 0.1179305. This tells us that an increase of 1 $\mu g/m^3$ of PM2.5 in a given California census tract is associated with an estimated increase of 0.1179305 in the percentage of births in that tract with weight less than 2500g. The effect of `pm25` on `low_birth_weight` *is* significant at the 5% level.

6 (e) New air quality policy

```
avg_new_lbw <- mean(df$low_birth_weight, na.rm = TRUE) - 2 * beta1_hat

lbw_ci_low <- mean(df$low_birth_weight, na.rm = TRUE) - 2 * model$conf.high[[2]]
lbw_ci_high <- mean(df$low_birth_weight, na.rm = TRUE) - 2 * model$conf.low[[2]]
```

The predicted average value of `low_birth_weight` given a reduction of 2 $\mu g/m^3$ of PM2.5 is 4.7675111. The 95% confidence interval for this value is 4.7345692 to 4.800453, which means there is a 95% probability that this interval contains the true population estimate for `low_birth_weight` given a 2 $\mu g/m^3$ reduction in `pm25`