EDS241: Assignment 1

Jake Eisaguirre

01/17/2022

Read in, inspect data, and wrangle

```
data <- read_excel(here("data","CES4.xlsx"))

clean_data <- data %>%
  as.data.frame() %>%
  select(`Census Tract`, `Total Population`, `Low Birth Weight`, PM2.5, Poverty) %>%
  na.omit() %>%
  clean_names() %>%
  filter(!str_detect(low_birth_weight, "NA"))
```

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

```
print(mean(clean_data$pm2_5))
```

```
## [1] 10.19529
```

The average concentration of PM2.5 across all census tracts in California is 10.15 micrograms per cubic meters

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

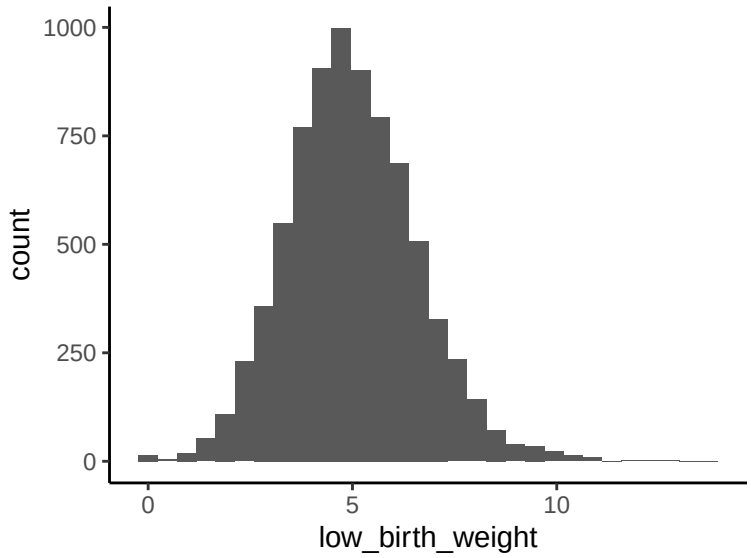
```
clean_data[which(clean_data$poverty == max(clean_data$poverty)), ]
```

```
##     census_tract total_population low_birth_weight    pm2_5 poverty
## 354   6037206300             6103            10.31 12.39953    93.2
```

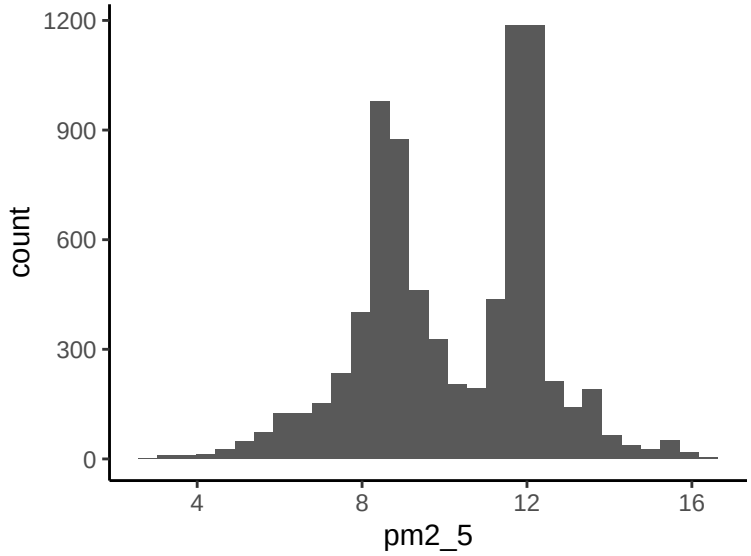
The county that has the highest level of poverty in California is census_tract: 6037206300

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

```
ggplot(data = clean_data, aes(x = as.numeric(low_birth_weight))) +
  geom_histogram() +
  theme_classic() +
  xlab("low_birth_weight")
```




```
ggplot(data = clean_data, aes(x = pm2_5)) +
  geom_histogram() +
  theme_classic()
```



(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 PM25 on LowBirthWeight statistically significant at the 5%?

```
model_1 <- lm_robust(low_birth_weight ~ pm2_5, data = clean_data)
summary(model_1)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5, data = clean_data)
##
```

```
## Standard error type:  HC2
##
## Coefficients:
##             Estimate Std. Error t value  Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)   3.7996   0.088578   42.90 0.000e+00   3.6259   3.9732 7803
## pm2_5         0.1182   0.008401   14.06 2.179e-44   0.1017   0.1346 7803
##
## Multiple R-squared:  0.02511 ,   Adjusted R-squared:  0.02499
## F-statistic: 197.8 on 1 and 7803 DF,  p-value: < 2.2e-16
```

The estimated slope coefficient is 0.1182 and its heteroskedasticity-robust standard error is 0.008401.

Slope interpretation: For every one unit increase in PM2.5 we see an 0.1182 percent increase in census tract birth weights less than 2500g.

Yes the effect of PM2.5 on low birth weights is statistically significant at the 5% level since the p-values is extremely small (2.179e-44).

(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 LowBirthWeight and derive its 95% confidence interval. Interpret the 95% confidence interval.

```
lbw <- as.numeric(clean_data$low_birth_weight)

new_average <- (0.1182*(mean(lbw)+2) + 3.7996)
print(new_average)
```

```
## [1] 4.627504
```

```
CI_pos <- mean(lbw) + 1.96 * (sd(lbw)/sqrt(7805))
CI_neg <- mean(lbw) - 1.96 * (sd(lbw)/sqrt(7805))
```

The new average value of birth weights less than 2500g is 4.627% with a decrease of 2 micrograms per cubic meters of PM2.5. We 95% confident that the population of baby birth weights less than 2500g is between 4.969% and 5.039%.

(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 PM25, compared to the regression in (d). Explain.

```
model_2 <- lm_robust(low_birth_weight ~ pm2_5 + poverty, data = clean_data)
summary(model_2)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = clean_data)
##
## 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
```

Poverty coefficient Interpretation: When PM2.5 is held fixed, a one unit increase in poverty will result in a 0.02744 percent increase in census tract birth weights less than 2500g.

The coefficient of PM2.5 in this model compared to the previous model seems to have decreased by 50%. This happens most likely due to the two predictor variables being some what correlated or having co-linearity. The model is unsure what variable is explaining the variance.

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

```
linearHypothesis(model_2, c("pm2_5=0", "poverty=0"), white.adjust = "hc2")
```

```
## Linear hypothesis test
##
## Hypothesis:
## pm2_5 = 0
## poverty = 0
##
## Model 1: restricted model
## Model 2: low_birth_weight ~ pm2_5 + poverty
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   7804
## 2   7802  2 989.61  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis that the effect of PM2.5 is equal to the effect of Poverty