EDS241: Assignment 1

Peter Menzies

01/18/2022

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 $\mu g/m^3$

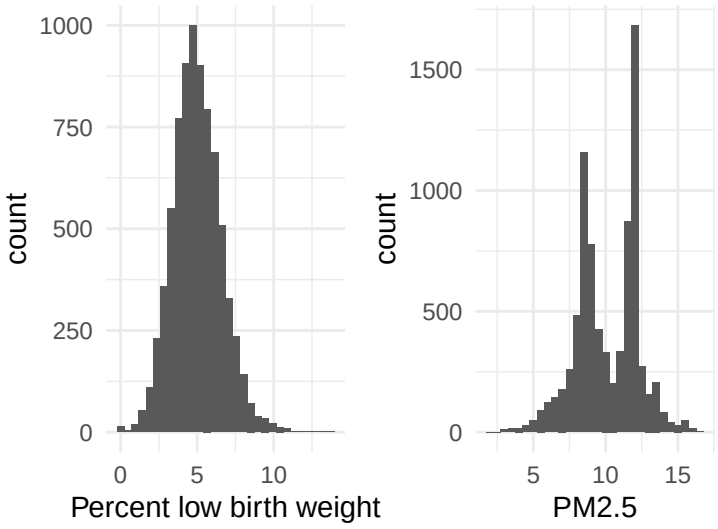
3 (b) What county has the highest level of poverty in California?

```
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_pm25 <- lm_robust(low_birth_weight ~ pm25, df)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	3.8010	0.0886	42.9088	0.0000	3.6273	3.9746	7806.0000	low_birth_weight
2	pm25	0.1179	0.0084	14.0354	0.0000	0.1015	0.1344	7806.0000	low_birth_weight

Table 1: Linear regression of percent low birth weight on PM2.5

```
beta1_hat <- model_pm25$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

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

lbw_ci_low <- mean(df$low_birth_weight, na.rm = TRUE) - 2 * model_pm25$conf.high[[2]]
lbw_ci_high <- mean(df$low_birth_weight, na.rm = TRUE) - 2 * model_pm25$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`

7 (f)

```r
model_pm25_poverty <- lm_robust(low_birth_weight ~ pm25 + poverty, df)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	3.5437	0.0847	41.8225	0.0000	3.3776	3.7098	7802.0000	low_birth_weight
2	pm25	0.0591	0.0083	7.1272	0.0000	0.0429	0.0754	7802.0000	low_birth_weight
3	poverty	0.0274	0.0010	27.3745	0.0000	0.0255	0.0294	7802.0000	low_birth_weight

Table 2: Multiple linear regression of percent low birth weight on PM2.5 and poverty

The estimated coefficient on `poverty` signifies that increasing the percentage of the population in the census tract living below twice the federal poverty line by 1% is associated with an estimated increase of 0.0274353 in `low_birth_weight`.

The estimated coefficient on `pm25` decreased from 0.1179305 in our original single variable model, to 0.0591077 in our multivariate model. This is because `pm25` and `poverty` are correlated (shown below)—and thus in our original model, the estimated coefficient on `pm25` was attempting to explain some of the variation in `low_birth_weight` that our second model shows to be associated with `poverty`.

```r
# showing correlation between pm25 and poverty
model_endogenous <- lm_robust(pm25 ~ poverty, df)
```

	term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
1	(Intercept)	9.2566	0.0428	216.2251	0.0000	9.1727	9.3405	7958.0000	pm25
2	poverty	0.0286	0.0013	22.4739	0.0000	0.0261	0.0311	7958.0000	pm25

Table 3: Linear regression of PM2.5 on poverty

8 (g) Linear hypothesis test

```r
lht <- linearHypothesis(model_pm25_poverty, "pm25=poverty", white.adjust = "hc2")
```

Based on the results of our linear hypothesis test, we can reject the null hypothesis that the effects of `pm25` and `poverty` on `low_birth_weight` are equal at the 0.1% level.

	res.df	df	statistic	p.value
1	7803.0000			
2	7802.0000	1.0000	13.4682	0.0002

Table 4: Comparing effects of PM2.5 and poverty on percent low birth weight