EDS241: HW1 solution key *(12 pts)*

Vincent Thivierge

01/26/2022

In this assignment, we use some of the underlying data of the CalEnviroScreen 4.0 to study the relationship between the percentage of low birth rates and ambient pollution in census tracts in California.

The following code chunk loads the data and converts our columns of interests in numeric format.

```r
# Load data

ces_raw <- read_excel("CES4.xlsx", sheet = 1)%>%
  clean_names()%>%
  as.data.table()

# Clean data

#Make sure relevant columns are numeric
ces_clean <- ces_raw %>%
  mutate_at(vars(low_birth_weight,pm2_5, poverty), as.numeric)
```

Question (a) Average PM$_{2.5}$ *(1 pt)*

The first question we can simply take an unweighted average of census tract level PM$_{2.5}$ concentrations

```r
mean_pm25 <- ces_clean%>%
  summarize(mean_pm25 = mean(pm2_5, na.rm=T))
```

The average PM$_{2.5}$ concentration in California census tracts is of 10.15 ug per cubic meter.

Question (b) County poverty rate *(1 pt)*

Since we observe both the poverty percentage and total population per census tract, we can find the county poverty rate by dividing the county level poverty rate over the total county level population. This is the population *weighted* average poverty rate.

```r
highest_poverty_cty <- ces_clean%>%
  group_by(california_county)%>%
  mutate(nb_poverty = (poverty/100)*total_population)%>%
  summarize(nb_poverty_cty = sum(nb_poverty, na.rm = T),
            pop_cty = sum(total_population, na.rm = T))%>%
  mutate(poverty_cty = nb_poverty_cty/pop_cty)%>%
  arrange(-poverty_cty)%>%
  slice_head(n=1)
```

The county with the highest poverty rate in California is Tulare County.

Question (c)   Histograms *(1 pt for each histogram)*

Figure 1 and 2 show the distribution of low birth rate percentages and $PM_{2.5}$ concentration in California census tracts. Figure 1 shows that low birth rates appear normally distributed, whereas $PM_{2.5}$ is moreso bimodaly distributed.

```
hist1 <- ces_clean%>%
ggplot(aes(low_birth_weight))+
  geom_histogram()+
  theme_cowplot(12)+
  labs(x = "Percentage of low birth weight rate", y = "Number of census tracts")

hist2 <- ces_clean%>%
  ggplot(aes(pm2_5))+
  geom_histogram()+
  theme_cowplot(12)+
  labs(x = "Annual mean PM2.5 concentrations", y = "Number of census tracts")
```

Figure 1: Distribution of census tract low birth weight rates

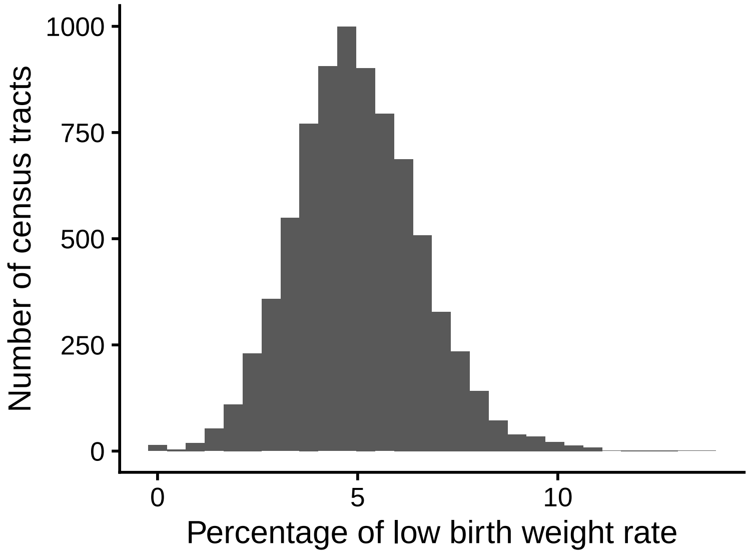
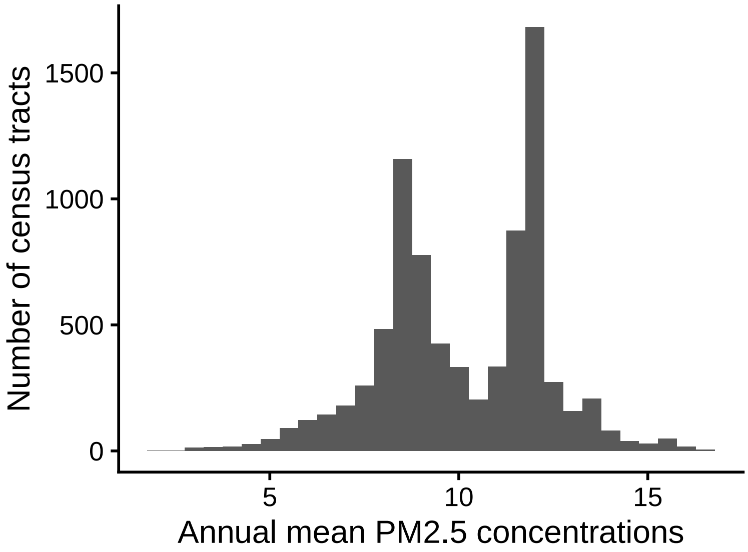


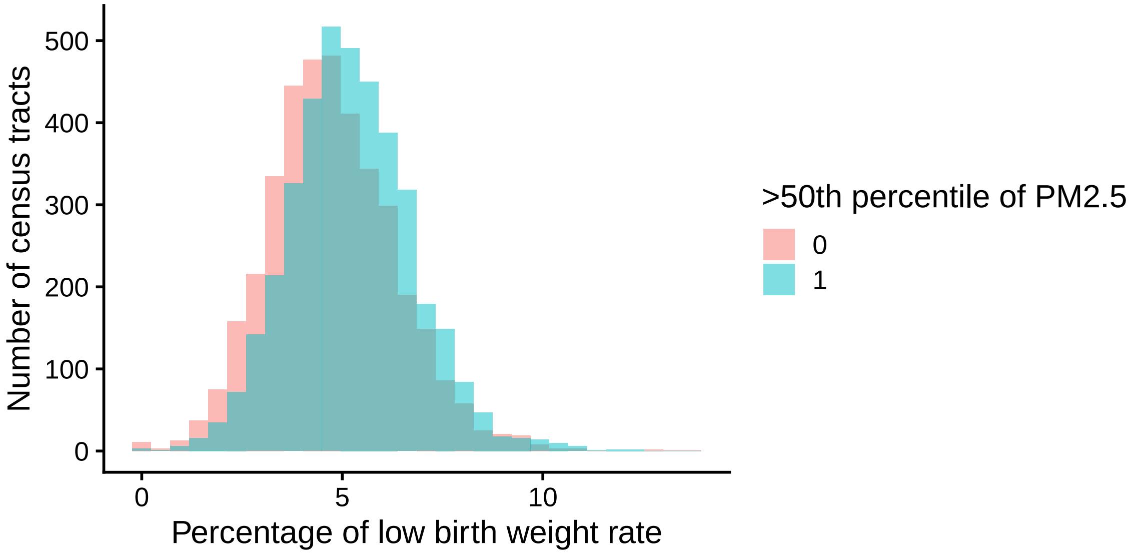
Figure 2: Distribution of census tract pm2.5 concentration



As an extra Figure, Figure 3 shows the conditional distribution of low birth weight rates for the lowest and highest $PM_{2.5}$ census tracts. As expected, the distribution of low birth rates for higher polluted census tracts is to the right of the lowest polluted tracts.

```
hist3 <- ces_clean%>%
  mutate(pm25_high = ifelse(pm2_5_pctl>50,1,0))%>%
  ggplot(aes(low_birth_weight, fill = as.factor(pm25_high)))+
  geom_histogram(alpha = 0.5, position = "identity")+
  theme_cowplot(12)+
  labs(x = "Percentage of low birth weight rate", y = "Number of census tracts",
       fill = ">50th percentile of PM2.5")
```

Figure 3: Conditional distributions of census tract low birth rates by pm2.5 concentration



Question (d) Univariate regression *(1 pt for estimation, 1 pt for standard error, 1 pt for interpretation)*

Table 1 show the estimate coefficient of regressing census tract low birth weight rates on PM2.5 concentrations. Our standard errors also account for heteroskedasticity.

```r
#With lm + estimar + stargazer

model1 <- lm(low_birth_weight ~ pm2_5, ces_clean)

##starprep() calculates robust standard errors
##starprep(model1) would give the same results as these are all the default arguments

se_model1 = starprep(model1,  stat = c("std.error"), se_type = "HC2", alpha = 0.05)

stargazer(model1, se = se_model1,
          type = "latex", ci=FALSE, no.space = TRUE,
          header = FALSE, omit = c("Constant"), omit.stat = c("adj.rsq","ser", "f"),
          covariate.labels = c("PM2.5 concentrations"), dep.var.labels = c("Percent low birth weight"),
          dep.var.caption = c(""),
          title = "Low birth weight and air pollution", table.placement = "H",
          notes = "Robust standard errors in parantheses", notes.align = "l")
```

Table 1: Low birth weight and air pollution

	Percent low birth weight
PM2.5 concentrations	0.118***
	(0.008)
Observations	7,808
R^2	0.025
Note:	*p<0.1; **p<0.05; ***p<0.01
	Robust standard errors in parantheses

The results suggest that a 1 ug per cubic meter increase in PM2.5 concentrations is associated with a 0.12 percentage point increase in low birth weight rate for the average California census tract. This relationship is also statistically different than zero.

Question (e)

(skipped)

Question (f) Bivariate regression *(1 pt for estimation, 1 pt for standard errors, 1 pt for interpretation)*

Table 2 additionally shows the effect of both PM2.5 and poverty rate on low birth weight rates. Both PM2.5 and poverty rate are positively and significantly linked with low birth weight rates in California. Adding the poverty rate decreases the coefficient on PM2.5 by about two times.

We can therefore think of poverty rate as an omitted variable in our initial model. Since the effect of poverty rate on low birth rate is positive, AND poverty rate and PM2.5 are positively correlated, the omission of poverty rate biased the PM2.5 coefficient upward.

```
model2 <- lm(low_birth_weight ~ pm2_5 + poverty, ces_clean)

se_models = starprep(model1, model2,  stat = c("std.error"), se_type = "HC2", alpha = 0.05)

stargazer(model1,model2, se = se_models,
          type = "latex", ci=FALSE, no.space = TRUE,
          header = FALSE, omit = c("Constant"), omit.stat = c("adj.rsq","ser", "f"),
          covariate.labels = c("PM2.5 concentrations", "Percent living below two times the federal pove
          dep.var.caption = c(""),
          title = "Low birth weight and air pollution", table.placement = "H",
          notes = "Robust standard errors in parantheses", notes.align = "l")
```

Table 2: Low birth weight and air pollution

	Percent low birth weight	
	(1)	(2)
PM2.5 concentrations	0.118***	0.059***
	(0.008)	(0.008)
Percent living below two times the federal poverty level		0.027***
		(0.001)
Observations	7,808	7,805
R^2	0.025	0.117
Note:	*p<0.1; **p<0.05; ***p<0.01 Robust standard errors in parantheses	

Question (g) Linear hypothesis test *(1 pt for F-test and 1 pt for interpretation)*

```
lin_test <- linearHypothesis(model2,c("pm2_5=poverty"), white.adjust = "hc2")
```

For the last section, we test whether the coefficients on PM2.5 and poverty rate are the same. This is equivalent to testing whether the difference between the coeffcients is equal to zero. With a P-value of our F-test of 0.0002443‘, we reject at the 5% level and even at the 0.01% level the null hypothesis is true, i.e. that the coefficients are equal.