

EDS241: Assignment 1

Erica Bishop

02/26/2023

```
#load libraries
library(tidyverse)
```

```
## — Attaching packages — tidyverse 1.3.2 —
## ✓ ggplot2 3.3.6      ✓ purrr 0.3.4
## ✓ tibble 3.1.8       ✓ dplyr 1.0.9
## ✓ tidyr 1.2.0        ✓ stringr 1.4.1
## ✓ readr 2.1.2        ✓ forcats 0.5.1
## — Conflicts — tidyverse_conflicts() —
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag() masks stats::lag()
```

```
library(estimatr)
library(stargazer)
```

```
##
## Please cite as:
##
## Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Ta
bles.
## R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(here)
```

```
## here() starts at /Users/ericabishop/Documents/MEDSwinter/EDS241-policy/eds241-policy-
evaluation
```

```
library(naniar)
library(patchwork)
```

```
#load data

CES_dat <- readxl::read_xlsx(here("eds241_data/CES4.xlsx")) |> #read-in file
  clean_names() #standardize variable names

#skimr::skim(CES_dat) #see what data looks like to clean

#clean data
CES_dat <- CES_dat |>
  select(census_tract, total_population, low_birth_weight, #select variables of interest
for this assignment
  pm2_5, poverty, linguistic_isolation) |>
  replace_with_na_all(condition = ~.x %in% "NA") |>
  mutate(low_birth_weight = as.numeric(low_birth_weight),
         linguistic_isolation = as.numeric(linguistic_isolation))
```

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

```
avg_pm2_5 <- mean(CES_dat$pm2_5, na.rm = TRUE)

print(paste0("The average PM 2.5 concentration across all census tracts in California is
", round(avg_pm2_5, 3), "."))
```

```
## [1] "The average PM 2.5 concentration across all census tracts in California is 10.15
3."
```

b. Make a histogram depicting the distribution of percent low birth weight and PM2.5.

```
hist_plot <- ggplot(data = CES_dat,
                   aes(fill = c(low_birth_weight, pm2_5))) +
  geom_histogram(aes(x = low_birth_weight,
                    fill = "Low Birth Weight"),
                col = "black",
                alpha = 0.5) +
  geom_histogram(aes(x = pm2_5,
                    fill = "PM 2.5"),
                col = "black",
                alpha = 0.5) +
  scale_fill_manual(values = c("blue", "yellow"),
                    name = "Variable") +
  labs(title = "Distrubition of low birth weight %s and PM 2.5 in California census trac
ts") +
  theme_minimal() +
  theme(
    axis.title.x = element_blank(),
  )

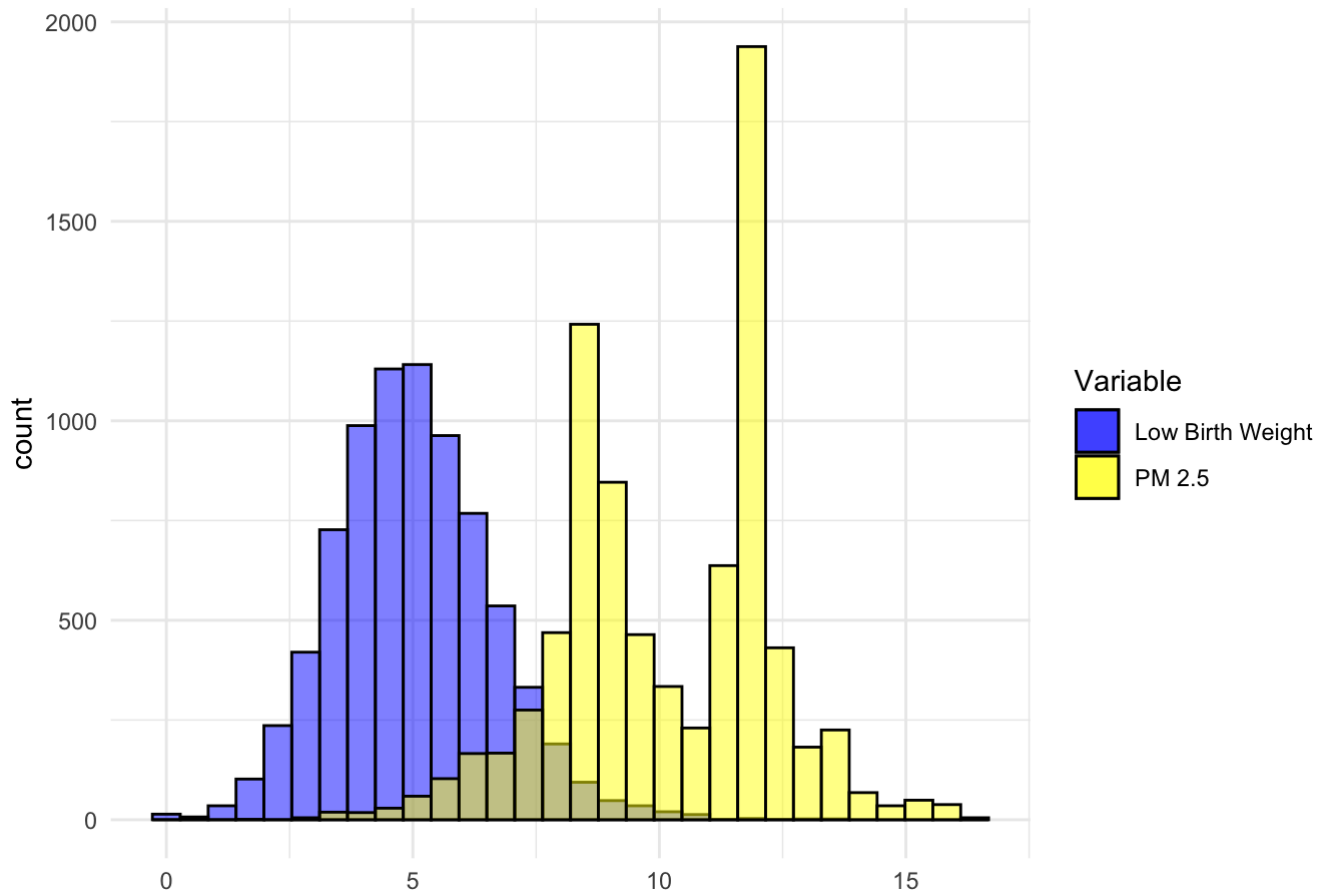
hist_plot
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 227 rows containing non-finite values (stat_bin).
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Distrubution of low birth weight %s and PM 2.5 in California census tracts



- c. Estimate an 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% level?

```
model1 <- lm(formula = low_birth_weight ~ pm2_5,
              data = CES_dat) #build model

#specify heteroskedasticic standard error
se_model1 <- stargazer(model1,
                       stat = c("std.error"),
                       se_type = "HC1",
                       alpha = 0.05)

stargazer(model1, se = se_model1, type="text")
```

```
##
## =====
##                               Dependent variable:
##                               -----
##                               low_birth_weight
## -----
## pm2_5                        0.118***
##                               (0.008)
##
## Constant                     3.801***
##                               (0.089)
##
## -----
## Observations                 7,808
## R2                          0.025
## Adjusted R2                 0.025
## Residual Std. Error         1.569 (df = 7806)
## F Statistic                 200.060*** (df = 1; 7806)
## =====
## Note:                        *p<0.1; **p<0.05; ***p<0.01
```

```
slope_coef <- model1$coefficients[2]
```

```
#print answers
```

```
print(paste("This slope coefficient from this model shows that there will be about an",
round(slope_coef, 3), "increase in low birth weight for a one unit increase in pm 2.5. T
his is statistically significant at the 5% level and the 1% level."))
```

```
## [1] "This slope coefficient from this model shows that there will be about an 0.118 i
ncrease in low birth weight for a one unit increase in pm 2.5. This is statistically sig
nificant at the 5% level and the 1% level."
```

```
print("The heteroskedastic robust standard error for the slope coefficient is The hetero
skedastic robust standard error for the slope coefficient is 0.008401 (this is very clos
e to the homoskedastic standard error of 0.00833).")
```

```
## [1] "The heteroskedastic robust standard error for the slope coefficient is The heter
oskedastic robust standard error for the slope coefficient is 0.008401 (this is very clo
se to the homoskedastic standard error of 0.00833)."
```

- 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.]

```

#create new df with lower PM2.5
CES_lowpm <- CES_dat |>
  mutate(pm2_5 = pm2_5 - 2)

#create new robust model
model2 <- lm_robust(low_birth_weight ~ pm2_5,
                    data = CES_dat, #use original df
                    se_type = "HC1",
                    alpha = 0.05)

#predict new birth weight
pred_lbwt <- predict(
  model2,
  newdata = CES_lowpm,
  se.fit = TRUE,
  interval = 'confidence'
)

#calculate fit average (center of confidence interval)
fit_avg_lbwt <- mean(pred_lbwt$fit)

original_lbwt <- mean(CES_dat$low_birth_weight, na.rm = TRUE)

#calculate averages for upper and lower bounds of confidence interval
low_lbwt <- mean(pred_lbwt$fit[,2])
high_lbwt <- mean(pred_lbwt$fit[,3])

#print answers
print(paste0("The new average low birth weight will likely be about ", round(fit_avg_lbwt, 3), "% of babies across census tracts in California if PM 2.5 decreases by 2 micrograms per cubic meter. "))

```

```
## [1] "The new average low birth weight will likely be about 4.762% of babies across census tracts in California if PM 2.5 decreases by 2 micrograms per cubic meter."
```

```
print(paste0("This is lower than the initial low birth weight of ", round(original_lbwt, 3), "% of babies across census tracts in California. "))
```

```
## [1] "This is lower than the initial low birth weight of 5.003% of babies across census tracts in California."
```

```
print(paste0("The interval from ", round(low_lbwt, 3), "to ", round(high_lbwt, 3), " will contain the true mean value of babies born with low birth rate across California Census tracts 95% of the time in this new scenario. "))
```

```
## [1] "The interval from 4.706 to 4.819 will contain the true mean value of babies born
with low birth rate across California Census tracts 95% of the time in this new scenario."
```

- 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.

```
model3 <- lm_robust(
  formula = low_birth_weight ~ pm2_5 + poverty,
  data = CES_dat,
  se_type = "HC1",
  alpha = 0.05
)

summary(model3)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5 + poverty, data = CES_dat,
##          se_type = "HC1", alpha = 0.05)
##
## Standard error type: HC1
##
## Coefficients:
##              Estimate Std. Error t value    Pr(>|t|) CI Lower CI Upper   DF
## (Intercept)  3.54374    0.084724  41.827 0.000e+00  3.37766  3.70982 7802
## pm2_5        0.05911    0.008292   7.128 1.108e-12  0.04285  0.07536 7802
## poverty      0.02744    0.001002  27.378 1.183e-157  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: < 2.2e-16
```

```
pov_coef <- model3$coefficients[3]
```

```
print(paste("The poverty coefficient means that low birth rates will increase by an additional", round(pov_coef, 3), "for every one unit increase in poverty."))
```

```
## [1] "The poverty coefficient means that low birth rates will increase by an additional 0.027 for every one unit increase in poverty."
```

```
print("The pm 2.5 coefficient in this new model, 0.059, is lower than the previous coefficient of 0.118. This is because poverty affects the birth rate so when it's accounted for in the model, the effect of pm 2.5 is smaller. The previous model incorrectly attributed all of the change in low birth weight to pm 2.5.")
```

```
## [1] "The pm 2.5 coefficient in this new model, 0.059, is lower than the previous coefficient of 0.118. This is because poverty affects the birth rate so when it's accounted for in the model, the effect of pm 2.5 is smaller. The previous model incorrectly attributed all of the change in low birth weight to pm 2.5."
```

- f. Create an indicator variable equal to 1 if the census tract is above the median Linguistic Isolation (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.

```
#create new df with indicator variable
CES_ling <- CES_dat |>
  add_column(ling_iso_threshold = case_when(
    CES_dat$linguistic_isolation > 6.9 ~ 1,
    TRUE ~ 0
  ))

#create new model
model4 <- lm_robust(
  formula = low_birth_weight ~ pm2_5 + poverty + ling_iso_threshold,
  data = CES_ling,
  se_type = "HC1",
  alpha = 0.05
)

summary(model4)
```

```
##
## Call:
## lm_robust(formula = low_birth_weight ~ pm2_5 + poverty + ling_iso_threshold,
##           data = CES_ling, se_type = "HC1", alpha = 0.05)
##
## Standard error type: HC1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|) CI Lower CI Upper  DF
## (Intercept)      3.62056   0.084909  42.640 0.000e+00  3.45411  3.78700 7801
## pm2_5            0.04879   0.008368   5.830 5.762e-09  0.03238  0.06519 7801
## poverty          0.02403   0.001165  20.626 4.403e-92  0.02175  0.02632 7801
## ling_iso_threshold 0.27650   0.040649   6.802 1.106e-11  0.19682  0.35619 7801
##
## Multiple R-squared:  0.1225 ,    Adjusted R-squared:  0.1222
## F-statistic: 360.4 on 3 and 7801 DF,  p-value: < 2.2e-16
```

```
ling_coef <- model4$coefficients[4]

print(paste("The coefficient on the linguistic isolation indicator variable is", round(ling_coef, 3), "which means above median linguistic isolation increases low birth weights by", round(ling_coef, 3)))
```

```
## [1] "The coefficient on the linguistic isolation indicator variable is 0.277 which means above median linguistic isolation increases low birth weights by 0.277"
```