R_Code_Sample

Xinghuan Luo

6/21/2021

I loaded required packages.

```
pkg_list <- c("tidyverse", "assertthat", "here", "stargazer")
lapply(pkg_list, require, character.only = TRUE)
```

I created different folders to store different results.

```
folder_path <- c("data", "documentation", "proc", "results", "scripts", "latex")

for (folder in folder_path) {
  suppressWarnings(dir.create(here(folder)))
}

results_path <- c("figures", "tables")

for (folder in results_path) {
  suppressWarnings(dir.create(here("results", folder)))
}
```

I imported two datasets, car_data and market_data.

```
initial_car_data<-
    list.files(path = here("data", "car_data"),
               pattern = "*.csv",
               full.names = TRUE) %>%
    map_df(~read_csv(., col_types = cols(.default = "c")))


initial_market_data<-
    list.files(path = here("data", "market_data"),
               pattern = "*.csv",
               full.names = TRUE) %>%
    map_df(~read_csv(., col_types = cols(.default = "c")))
```

I merged the two datasets together and fill in the missing values.

```
initial_market_data <- initial_market_data %>%
  mutate(ma_key = case_when(
    ma == "B" ~ "Belgium",
    ma == "F" ~ "France",
    ma == "G" ~ "Germany",
    ma == "I" ~ "Italy",
    ma == "U" ~ "UK"
    ),
    ye = as.numeric(ye)
```

```
  )

initial_car_data <-  rename(initial_car_data, ma_key = ma) %>%
  mutate(ye = as.numeric(ye) + 1900)

to_numeric <- c("li1", "li2", "li3", "li", "hp", "qu")
final_data <- inner_join(
  initial_car_data,
  initial_market_data,
  by = c("ma_key", "ye")
  ) %>%
  mutate(
    across(
      all_of(to_numeric),
      as.numeric)
  ) %>%
  as_tibble()

nonmiss_avg <- function(data, miss_var, var1, var2, var_avg) {
  data %>%
    mutate({{miss_var}} := if_else(is.na({{miss_var}}),
                            3* {{var_avg}}-({{var1}} + {{var2}}),
                            {{miss_var}}))
}

final_data <- final_data %>%
  nonmiss_avg(li3, li1, li2, li) %>%
  nonmiss_avg(li2, li3, li1, li) %>%
  nonmiss_avg(li1, li3, li2, li) %>%
  mutate(li = if_else(is.na(li), (li1 + li2 + li3)/3, li))

non_missing <-  final_data %>%
  filter(is.na(li1) | is.na(li2) |is.na(li3) | is.na(li)) %>%
  nrow == 0

assert_that(non_missing)
```

I generated the variables for graph and regressions.

```
final_data_7090 <- final_data %>%
  filter(ye == 1970 | ye == 1990) %>%
  group_by(ye) %>%
  mutate(decile = ntile(hp, 10)) %>%
  group_by(ye, decile) %>%
  mutate(li_mean = weighted.mean(li, qu),
         mid_hp = median(hp),
         li_mean = mean(li_mean),
         ye = as.character(ye),
         decile = as.numeric(decile),
         log_hp = log(hp))

reg_70 = lm(li ~ hp + log_hp, weights = qu, data = subset(final_data_7090, ye == "1970"))
reg_90 = lm(li ~ hp + log_hp, weights = qu, data = subset(final_data_7090, ye == "1990"))
```

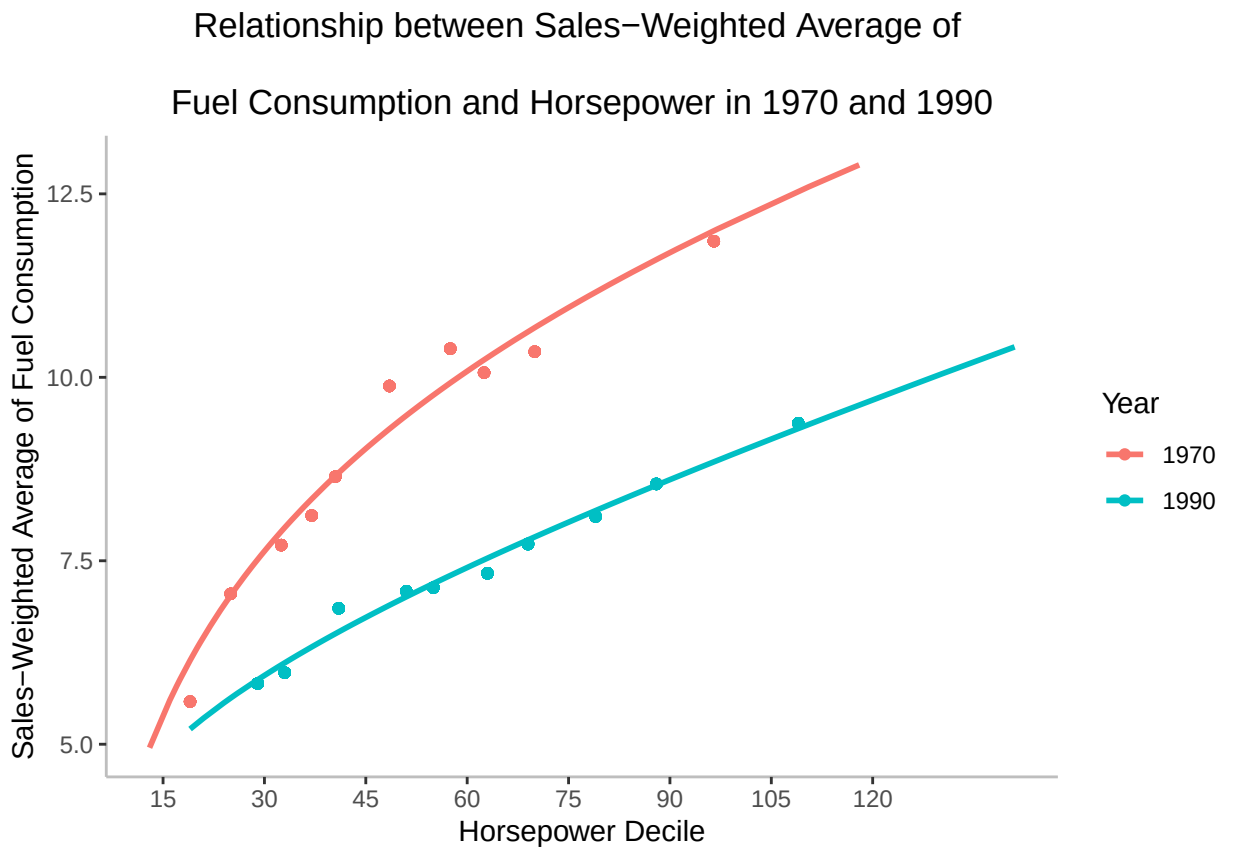
```
predicted_70 <- predict(reg_70, interval = "confidence") %>%
  as_tibble()
predicted_90 <- predict(reg_90, interval = "confidence") %>%
  as_tibble()

predicted_7090 <- bind_rows(predicted_70, predicted_90)

final_data_7090 <-  cbind(final_data_7090, predicted_7090) %>%
  arrange(hp)
```

I generated the graphs.

```
ggplot(final_data_7090, aes(x = mid_hp, y = li_mean,  color = ye)) +
  geom_point() +
  geom_line(aes(x = hp , y = fit, color = ye), size = 1) +
  labs(title = "Relationship between Sales-Weighted Average of
                \nFuel Consumption and Horsepower in 1970 and 1990",
       x = "Horsepower Decile",
       y = "Sales-Weighted Average of Fuel Consumption",
       color = "Year") +
  scale_x_continuous(breaks = seq(0, 120, 15)) +
  theme(
    plot.title = element_text(hjust = 0.5),
    panel.background = element_blank(),
    axis.line = element_line(colour = "grey"),
    legend.key = element_rect(fill = "white") ,
  )
```




```
ggsave("scatter_fitted.png",
       width = 14,
       height = 7,
       path = here("results", "figures"))
```

The scatterplot plots the sales-weighted average of fuel consumption versus the midpointof each horsepower decile. The fitted curves regress fuel consumption on horsepower and log horsepower, using sales as sample weights.

I generated the regression tables.

```
stargazer(reg_70, reg_90,
          title = "Regression Results",
          column.labels =  c("1970", "1990"),
          column.separate = c(1, 1),
          dep.var.labels = "Fuel Consumption",
          covariate.labels = c("Horspower", "Log Horsepower"),
          omit.stat = c("adj.rsq", "res.dev"),
          df = FALSE,
          colnames = FALSE,
          model.numbers = FALSE,
          out = here("results", "tables", "regression_results.tex")
  )
```

Table 1: Regression Results

	Dependent variable:	
	Fuel Consumption	
	1970	1990
Horspower	0.015	0.027***
	(0.011)	(0.008)
Log Horsepower	2.908***	0.947**
	(0.407)	(0.422)
Constant	−2.696**	1.904
	(1.064)	(1.247)
Observations	272	398
R^2	0.753	0.658
Residual Std. Error	129.530	92.511
F Statistic	410.846***	380.079***
Note:	*p<0.1; **p<0.05; ***p<0.01	