rajan_hwk3_s1

2025-02-24

```r
if (!require("pacman")) install.packages("pacman")
```

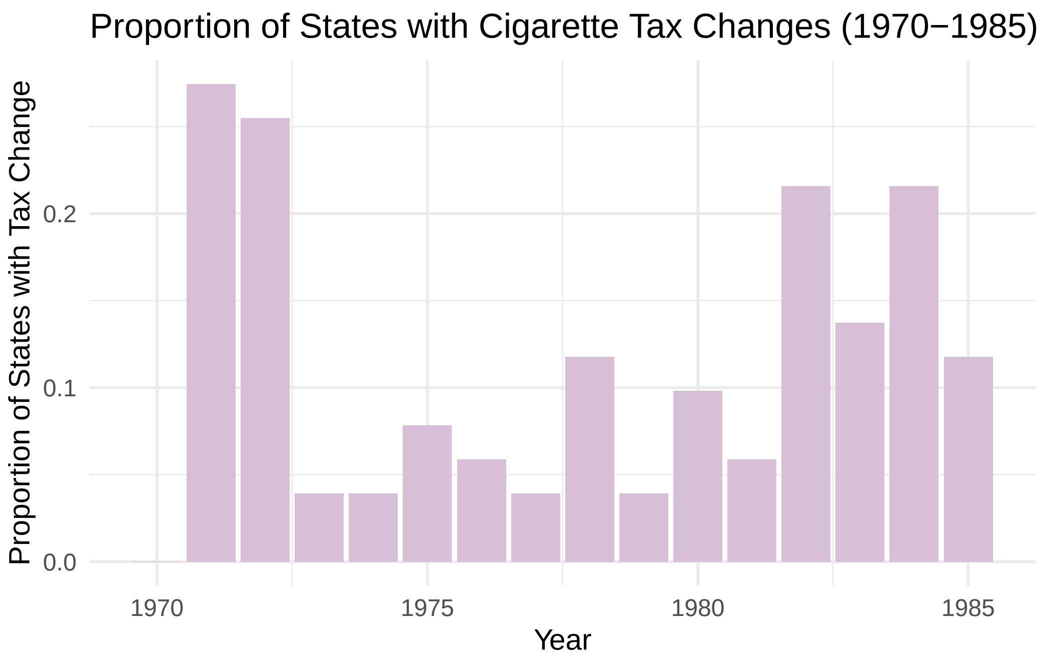
```
Loading required package: pacman
```

```r
pacman::p_load(tidyverse, ggplot2, dplyr, lubridate, stringr, readxl, data.table, gdata, fixe

load("/Users/sushmitarajan/econ470spring2025/Homework3/submission1/results/Hwk3_workspace.RDa
```

Click here to view my repository 1. Present a bar graph showing the proportion of states with a change in their cigarette tax in each year from 1970 to 1985.

```r
# Summarize tax changes per state-year
tax_burden_final_changes <- tax_burden_final %>%
  arrange(state, Year) %>%
  group_by(state) %>%
  mutate(tax_change = ifelse(Year == 1970, FALSE,
                             ifelse(is.na(lag(tax_state)) | tax_state != lag(tax_state), TRU
  ungroup()

# Calculate the proportion of states with tax changes each year
proportion_changes <- tax_burden_final_changes %>%
  group_by(Year) %>%
  summarize(proportion_changed = mean(tax_change, na.rm = TRUE)) %>%
  filter(Year >= 1970 & Year <= 1985)

# Plot the bar graph
ggplot(proportion_changes, aes(x = Year, y = proportion_changed)) +
  geom_bar(stat = "identity", fill = "#D8BFD8") +
  labs(title = "Proportion of States with Cigarette Tax Changes (1970-1985)",
```

```
       x = "Year",
       y = "Proportion of States with Tax Change") +
  theme_minimal()
```

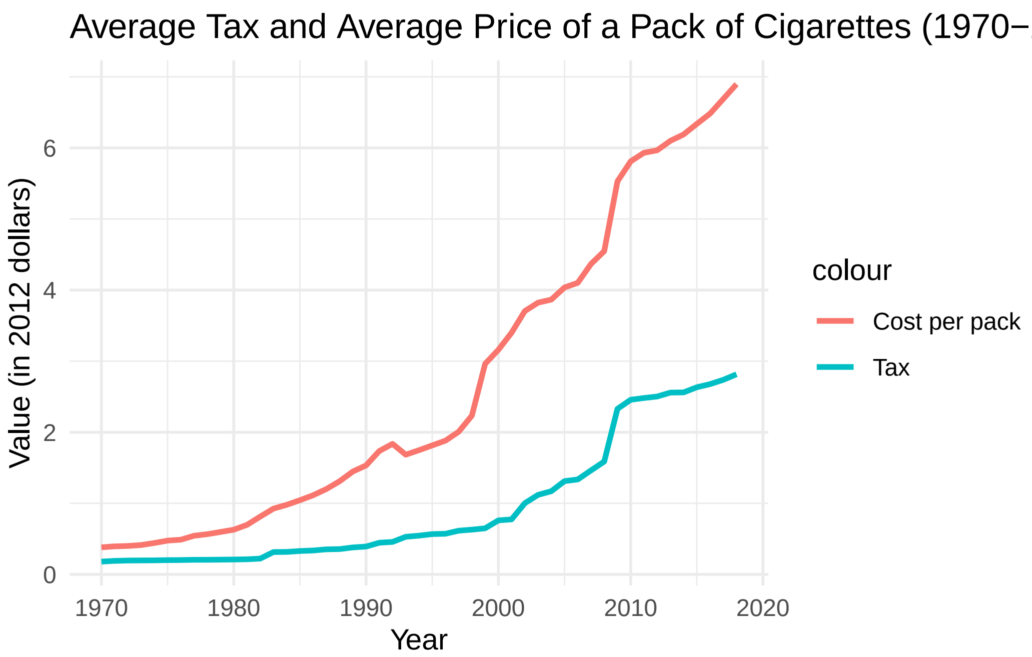


2. Plot on a single graph the average tax (in 2012 dollars) on cigarettes and the average price of a pack of cigarettes from 1970 to 2018.

```
# Aggregate the data by Year
aggregated_data <- tax_burden_final %>%
  group_by(Year) %>%
  summarise(
    avg_tax_dollar = mean(tax_dollar, na.rm = TRUE),     # Average tax in 2012 dollars
    avg_cost_per_pack = mean(cost_per_pack, na.rm = TRUE) # Average cost per pack
  )

# Create the plot with the aggregated data
ggplot(aggregated_data, aes(x = Year)) +
  geom_line(aes(y = avg_tax_dollar, color = "Tax"), size = 1) +
  geom_line(aes(y = avg_cost_per_pack, color = "Cost per pack"), size = 1) +
  labs(
    title = "Average Tax and Average Price of a Pack of Cigarettes (1970-2018)",
    x = "Year",
    y = "Value (in 2012 dollars)"
```

```
  ) +
  theme_minimal()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```

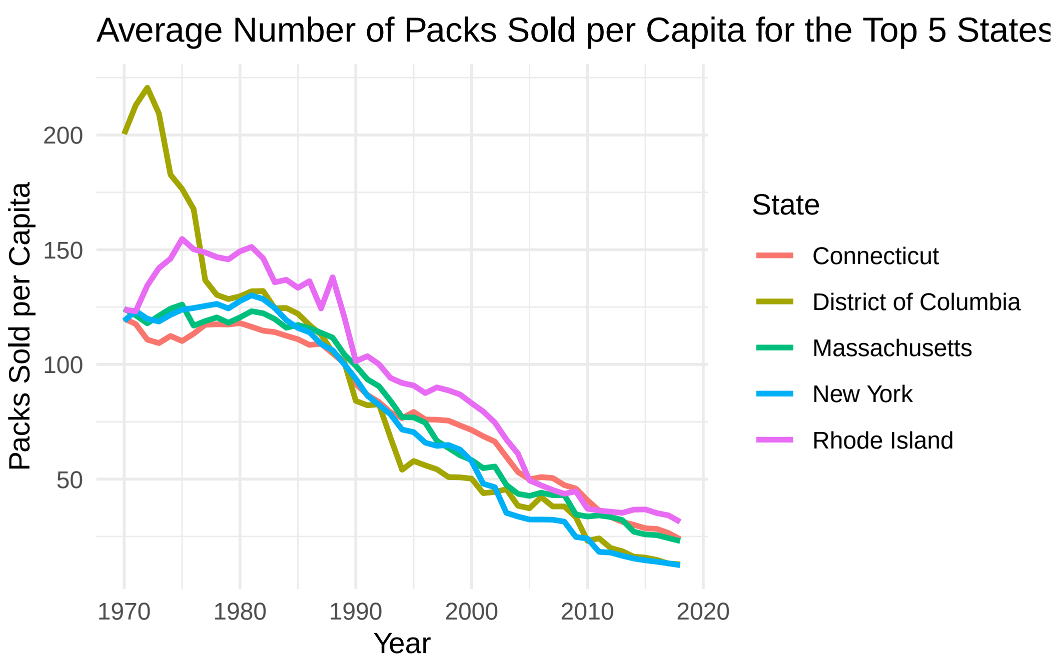


3. Identify the 5 states with the highest increases in cigarette prices (in dollars) over the time period. Plot the average number of packs sold per capita for those states from 1970 to 2018.

```
price_change <- tax_burden_final %>%
  group_by(state) %>%
  filter(Year == 1970 | Year == 2018) %>%
  summarise(price_change = cost_per_pack[Year == 2018] - cost_per_pack[Year == 1970])

top_states <- price_change %>%
  arrange(desc(price_change)) %>%
  head(5)

print(paste("States with the highest increases in cigarette prices:", paste(top_states$state
```

```
[1] "States with the highest increases in cigarette prices: New York, District of Columbia, C
```

```r
top_states_final <- tax_burden_final %>%
  filter(state %in% top_states$state)


ggplot(top_states_final, aes(x = Year, y = sales_per_capita, color = state)) +
  geom_line(size = 1) +
  labs(
    title = "Average Number of Packs Sold per Capita for the Top 5 States with the Highest Pr
    x = "Year",
    y = "Packs Sold per Capita",
    color = "State"
  ) +
  theme_minimal()
```

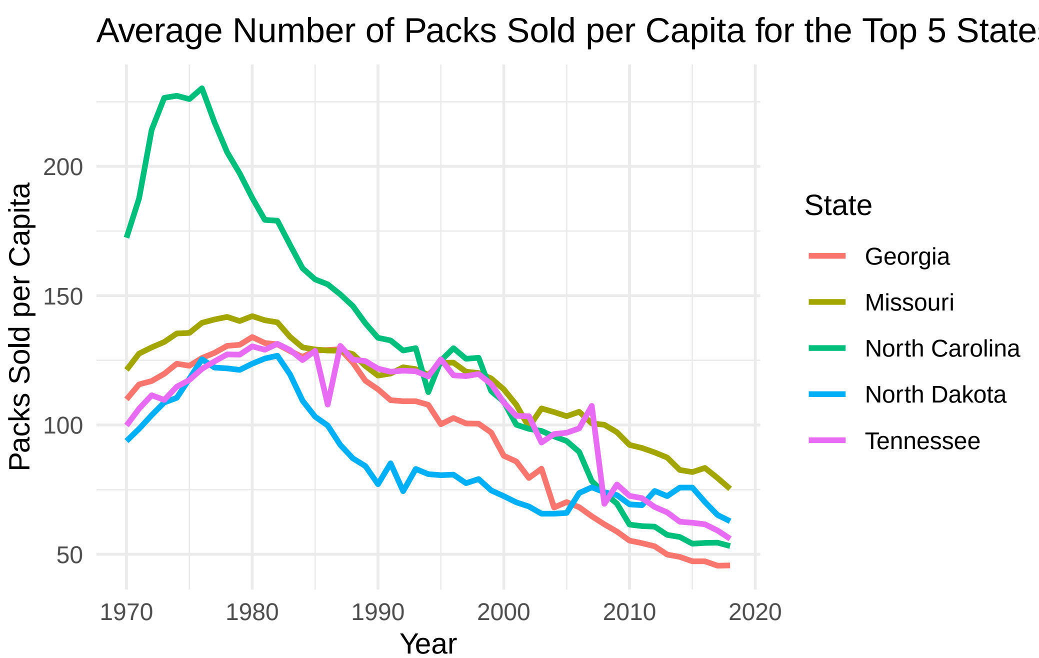


4. Identify the 5 states with the lowest increases in cigarette prices over the time period. Plot the average number of packs sold per capita for those states from 1970 to 2018.

```r
low_states <- price_change %>%
  arrange(price_change) %>%
  head(5)

print(paste("States with the lowest increases in cigarette prices:", paste(low_states$state,
```

```
[1] "States with the lowest increases in cigarette prices: Missouri, North Dakota, Tennessee
```

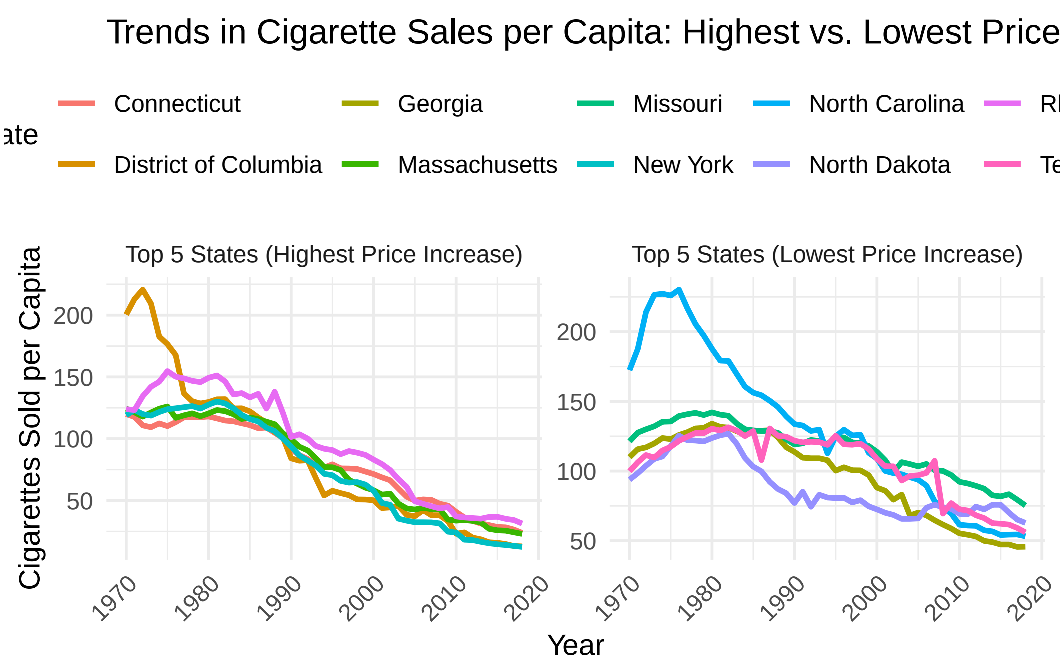
```
low_states_final <- tax_burden_final %>%
  filter(state %in% low_states$state)


ggplot(low_states_final, aes(x = Year, y = sales_per_capita, color = state)) +
  geom_line(size = 1) +
  labs(
    title = "Average Number of Packs Sold per Capita for the Top 5 States with the Highest Pr
    x = "Year",
    y = "Packs Sold per Capita",
    color = "State"
  ) +
  theme_minimal()
```



5. Compare the trends in sales from the 5 states with the highest price increases to those with the lowest price increases.

```
#Combine both datasets for plotting
high_low_combined <- bind_rows(
  top_states_final %>% mutate(group = "Top 5 States (Highest Price Increase)"),
  low_states_final %>% mutate(group = "Top 5 States (Lowest Price Increase)")
```

```
)

# Step 5: Plot the trends for sales_per_capita in both groups using ggplot
ggplot(high_low_combined, aes(x = Year, y = sales_per_capita, color = state)) +
  geom_line(size = 1) +
  facet_wrap(~group, scales = "free_y") +  # Separate the plots by group
  labs(
    title = "Trends in Cigarette Sales per Capita: Highest vs. Lowest Price Increases (1970-2
    x = "Year",
    y = "Cigarettes Sold per Capita",
    color = "State"
  ) +
  theme_minimal() +
  theme(legend.position = "top") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels
```




```
## The amount of cigarettes sold in both the highest and lowest states has drastically decre
```

6. Focusing only on the time period from 1970 to 1990, regress log sales on log prices to estimate the price elasticity of demand over that period. Interpret your results.

```r
# Filter data for the years 1970 to 1990 (if not already done)
tax_burden_1970_1990 <- tax_burden_final %>% filter(Year >= 1970 & Year <= 1990)

# Create log-transformed variables for sales and price
cig.data_1970_1990 <- tax_burden_1970_1990 %>% mutate(ln_sales=log(sales_per_capita),
                                ln_price_cpi=log(price_cpi),
                                ln_price=log(cost_per_pack),
                                tax_cpi=tax_state*(230/index),
                                total_tax_cpi=tax_dollar*(230/index),
                                ln_total_tax=log(total_tax_cpi),
                                ln_state_tax=log(tax_cpi))
# Run the regression
ols <- lm(ln_sales ~ ln_price, data=cig.data_1970_1990)


# Display the summary of the regression model
summary(ols)
```

```

Call:
lm(formula = ln_sales ~ ln_price, data = cig.data_1970_1990)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.77629 -0.09967 -0.00787  0.09969  0.78423 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.750402   0.008116   585.3   <2e-16 ***
ln_price    -0.171540   0.013829   -12.4   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2107 on 1069 degrees of freedom
Multiple R-squared:  0.1258,	Adjusted R-squared:  0.125 
F-statistic: 153.9 on 1 and 1069 DF,  p-value: < 2.2e-16
```

7. Again limiting to 1970 to 1990, regress log sales on log prices using the total (federal and state) cigarette tax (in dollars) as an instrument for log prices. Interpret your results and compare your estimates to those without an instrument. Are they different? If so, why?

```r
# Instrumental variable regression
ivs <- feols(ln_sales ~ 1 | ln_price ~ ln_total_tax, data = cig.data_1970_1990)

# Display the summary of the IV regression results
summary(ivs)
```

```
TSLS estimation - Dep. Var.: ln_sales
                  Endo.    : ln_price
                  Instr.   : ln_total_tax
Second stage: Dep. Var.: ln_sales
Observations: 1,071
Standard-errors: IID
             Estimate Std. Error   t value   Pr(>|t|)
(Intercept)  4.991108   0.034106 146.34225  < 2.2e-16 ***
fit_ln_price 0.502373   0.089837   5.59207 2.8482e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
RMSE: 0.377891   Adj. R2: -1.81869
F-test (1st stage), ln_price: stat =  88.4, p < 2.2e-16, on 1 and 1,069 DoF.
                  Wu-Hausman: stat = 240.2, p < 2.2e-16, on 1 and 1,068 DoF.
```

8. Show the first stage and reduced-form results from the instrument.

```r
# First Stage: Regression of ln_price on ln_total_tax
step1 <- lm(ln_price ~ ln_total_tax, data = cig.data_1970_1990)
summary(step1)
```

```
Call:
lm(formula = ln_price ~ ln_total_tax, data = cig.data_1970_1990)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0060 -0.3359 -0.1095  0.3779  1.0872

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.48152    0.01904 -25.296   <2e-16 ***
ln_total_tax -0.41181    0.04381  -9.399   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.4479 on 1069 degrees of freedom
Multiple R-squared:  0.07634,	Adjusted R-squared:  0.07547 
F-statistic: 88.35 on 1 and 1069 DF,  p-value: < 2.2e-16
```

```r
# Predicted values for ln_price
pricehat <- predict(step1)

# Reduced-form: Regression of ln_sales on predicted ln_price
step2 <- lm(ln_sales ~ pricehat, data = cig.data_1970_1990)
summary(step2)
```

```
Call:
lm(formula = ln_sales ~ pricehat, data = cig.data_1970_1990)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.86239 -0.09798  0.00549  0.09359  0.95094 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.99111    0.01947 256.365   <2e-16 ***
pricehat     0.50237    0.05128   9.796   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2159 on 1069 degrees of freedom
Multiple R-squared:  0.08238,	Adjusted R-squared:  0.08152 
F-statistic: 95.97 on 1 and 1069 DF,  p-value: < 2.2e-16
```

9. Repeat questions 1-3 focusing on the period from 1991 to 2015.

```r
# Filter data for 1991-2015
tax_burden_1991_2015 <- tax_burden_final %>% filter(Year >= 1991 & Year <= 2015)

# Log-transformed variables for sales and price
cig.data_1991_2015 <- tax_burden_1991_2015 %>% mutate(ln_sales=log(sales_per_capita),
                                ln_price_cpi=log(price_cpi),
                                ln_price=log(cost_per_pack),
                                tax_cpi=tax_state*(230/index),
                                total_tax_cpi=tax_dollar*(230/index),
```

```r
                                ln_total_tax=log(total_tax_cpi),
                                ln_state_tax=log(tax_cpi))

# Run the regression
ols <- lm(ln_sales ~ ln_price, data = cig.data_1991_2015)
summary(ols)
```

```
Call:
lm(formula = ln_sales ~ ln_price, data = cig.data_1991_2015)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.9375 -0.1781  0.0013  0.1860  1.1433 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.03949    0.02291  219.93   <2e-16 ***
ln_price    -0.66563    0.01747  -38.09   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3056 on 1273 degrees of freedom
Multiple R-squared:  0.5327,    Adjusted R-squared:  0.5323 
F-statistic:  1451 on 1 and 1273 DF,  p-value: < 2.2e-16
```

```r
# IV regression
ivs <- feols(ln_sales ~ 1 | ln_price ~ ln_total_tax, data = cig.data_1991_2015)
summary(ivs)
```

```
TSLS estimation - Dep. Var.: ln_sales
                  Endo.    : ln_price
                  Instr.   : ln_total_tax
Second stage: Dep. Var.: ln_sales
Observations: 1,275
Standard-errors: IID 
              Estimate Std. Error  t value  Pr(>|t|)    
(Intercept)   5.218896   0.026498 196.9549 < 2.2e-16 ***
fit_ln_price -0.813109   0.020548 -39.5712 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
RMSE: 0.313741   Adj. R2: 0.506163
F-test (1st stage), ln_price: stat = 4,111.4, p < 2.2e-16, on 1 and 1,273 DoF.
                  Wu-Hausman: stat =   280.6, p < 2.2e-16, on 1 and 1,272 DoF.
```

```r
# First Stage Regression
step1 <- lm(ln_price ~ ln_total_tax, data = cig.data_1991_2015)
summary(step1)
```

```
Call:
lm(formula = ln_price ~ ln_total_tax, data = cig.data_1991_2015)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.57946 -0.17145  0.02125  0.16675  0.66985 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.045904   0.007185  145.57   <2e-16 ***
ln_total_tax 0.726380   0.011328   64.12   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2383 on 1273 degrees of freedom
Multiple R-squared:  0.7636,    Adjusted R-squared:  0.7634 
F-statistic:  4111 on 1 and 1273 DF,  p-value: < 2.2e-16
```

```r
# Predict the fitted values for ln_price
pricehat <- predict(step1)

# Display the predicted values
cat("\nPredicted values of ln_price (pricehat) from the first stage regression:\n")
```

```
Predicted values of ln_price (pricehat) from the first stage regression:
```

```r
print(head(pricehat))
```

```
        1         2         3         4         5         6 
0.6944512 0.6727770 0.7271831 0.7084859 0.6883885 0.6674029 
```

```r
# Reduced-form Regression
step2 <- lm(ln_sales ~ pricehat, data = cig.data_1991_2015)
summary(step2)
```

```

Call:
lm(formula = ln_sales ~ pricehat, data = cig.data_1991_2015)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.90878 -0.15465  0.01119  0.15334  1.16925 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.21890    0.02365  220.69   <2e-16 ***
pricehat    -0.81311    0.01834  -44.34   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2802 on 1273 degrees of freedom
Multiple R-squared:  0.607,	Adjusted R-squared:  0.6067 
F-statistic:  1966 on 1 and 1273 DF,  p-value: < 2.2e-16
```

10. Compare your elasticity estimates from 1970-1990 versus those from 1991-2015. Are they different? If so, why?

1970-1990 shows a positive elasticity (i.e., price increase leads to more sales), which is unusual for most markets but could be explained by certain market factors like tax increases, changes in policy (e.g., tobacco taxation), or other structural shifts. For instance, if a tax increase made the product appear more "exclusive" or "prestigious," people might have bought more despite the higher price.

1991-2015 shows the more typical negative elasticity, where higher prices are associated with lower sales. This is consistent with standard economic theory and consumer behavior, where higher prices lead to a reduction in demand.

```r
elasticity_1970_1990 <- summary(ols)$coefficients["ln_price", "Estimate"]
elasticity_1991_2015 <- summary(ols)$coefficients["ln_price", "Estimate"]

cat("\nPrice Elasticity of Demand from 1970-1990: ", elasticity_1970_1990, "\n")
```

```

Price Elasticity of Demand from 1970-1990:  -0.6656264
```

```r
cat("Price Elasticity of Demand from 1991-2015: ", elasticity_1991_2015, "\n")
```

```
Price Elasticity of Demand from 1991-2015:  -0.6656264
```