Panel Data Analysis Using Cigarette Consumption Data

Devmini Wimlaasena, Manuela Lozano, Nicole Shock, Kyle Pasco

June 5, 2025

Contents

1 Part A

1.1 Introduction and Data Discussion

This project examines the impact of cigarette prices and consumer income on per capita cigarette consumption across U.S. states over time. The analysis employs a panel regression framework to address both cross-sectional (state-level) and temporal (yearly) variations, which allows for a more robust estimation of the determinants of cigarette consumption by controlling for unobserved heterogeneity.

The dataset utilized is the Cigar dataset from the Ecdat package in R. This dataset comprises a balanced panel of 1,020 observations, representing 46 U.S. states over a 30-year period from 1963 to 1992. The data were compiled from various U.S. government sources and are frequently used in applied econometrics research, particularly in studies related to health economics and taxation policy.

The key variables analyzed include:

- `sales`: Per capita sales of cigarettes (in packs)
- `price`: Average price per pack (in cents)
- `ndi`: Per capita income (in dollars)
- `pop`: State population size (in thousands)
- `year`: Year of observation
- `state`: State identifier

These variables provide a comprehensive view of both economic and demographic factors influencing cigarette consumption. The panel structure of the data enables the identification and control of state-specific and time-specific effects, which is crucial for isolating the causal impact of price and income on cigarette sales.

The primary research question guiding this analysis is: To what extent do cigarette prices and income levels influence per capita cigarette consumption across U.S. states over time?

This question is addressed through a sequence of descriptive and econometric analyses, as outlined in the project guidelines. The descriptive analysis provides an overview of the data and highlights potential heterogeneity across states and years, setting the stage for the subsequent panel regression modeling.

2 Part B: Descriptive Analysis

To better understand the nature of the dataset and identify underlying trends, we begin with descriptive statistics and graphical summaries of the main variables: `sales`, `price`, and `ndi`. These outputs help reveal potential individual- and time-level heterogeneity.

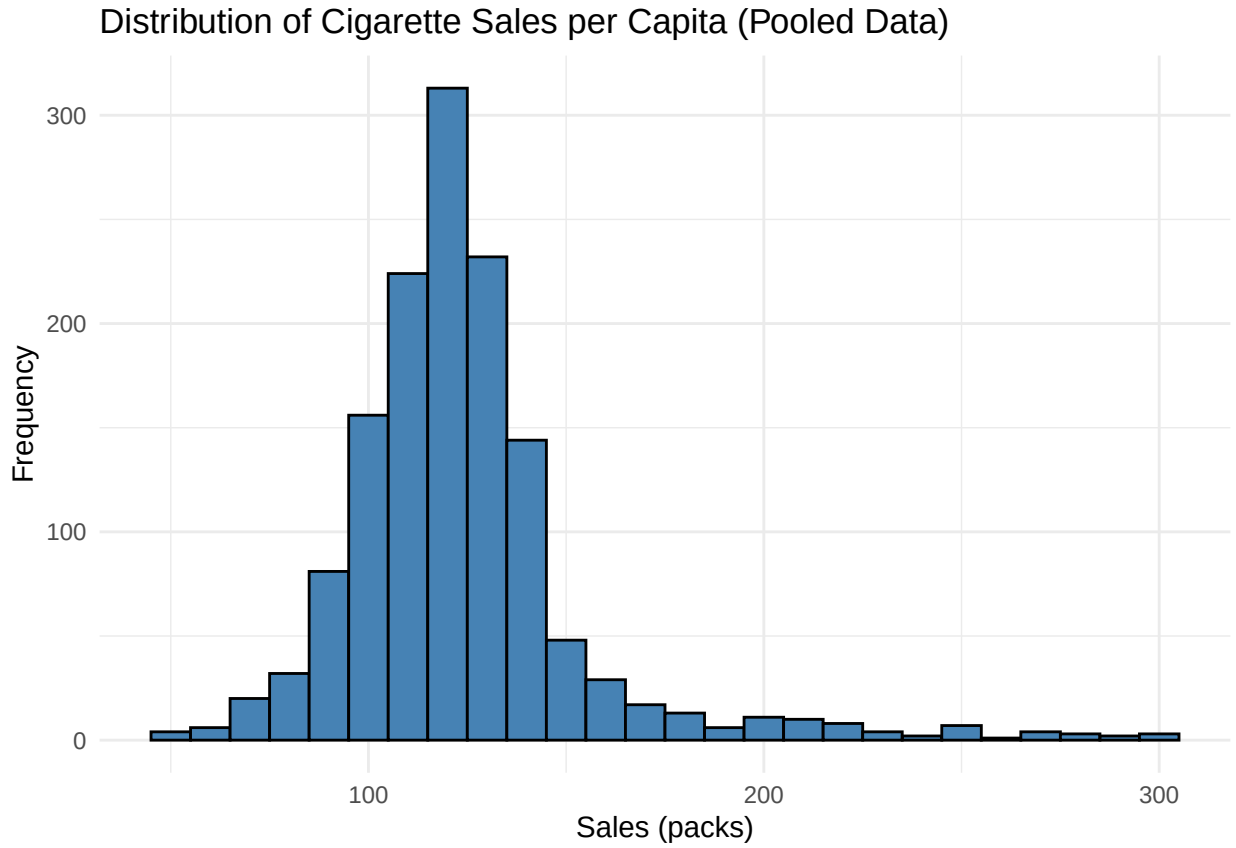
2.1 Summary Statistics

library(plm) library(lmtest) library(sandwich) library(stargazer)

```
##      sales            price             ndi
##  Min.   : 53.4   Min.   : 23.40   Min.   : 1323
##  1st Qu.:107.9   1st Qu.: 34.77   1st Qu.: 3328
##  Median :121.2   Median : 52.30   Median : 6281
##  Mean   :124.0   Mean   : 68.70   Mean   : 7525
##  3rd Qu.:133.2   3rd Qu.: 98.10   3rd Qu.:11024
##  Max.   :297.9   Max.   :201.90   Max.   :23074
```

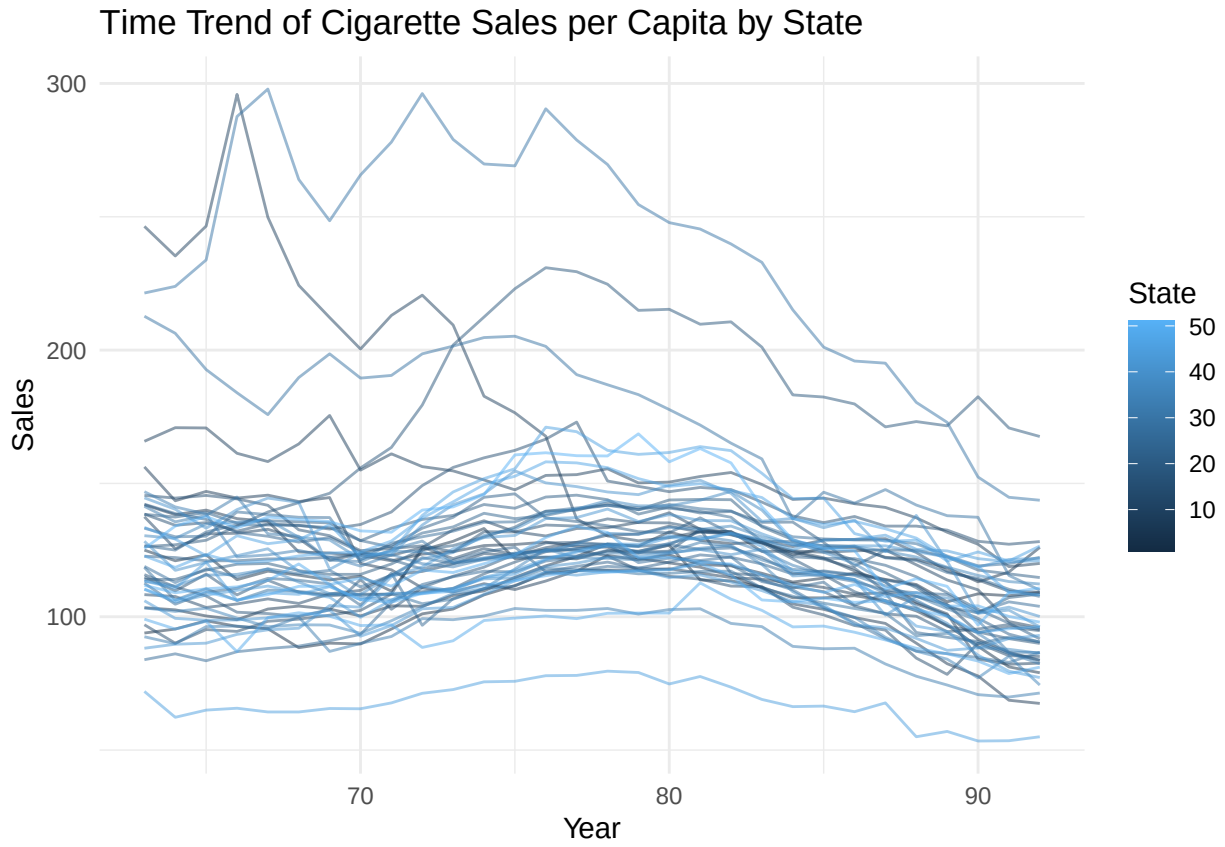
The summary statistics for the panel dataset—spanning 46 U.S. states over 30 years—demonstrate considerable variation in cigarette consumption, price, and income. Per capita cigarette sales range from 53.4 to 297.9 packs, with a mean of 124.0 packs. The average price per pack varies from 23.4 to 201.9 cents, averaging 68.7 cents, while per capita income ranges from $1,323 to $23,074, with a mean of $7,525. These ranges highlight substantial cross-sectional and temporal heterogeneity, underscoring the necessity of panel data methods to control for unobserved differences between states and over time. The dispersion in these variables also suggests the influence of state-specific factors and changing economic conditions on cigarette consumption.

2.2 Histogram of Cigarette Consumption (Sales)



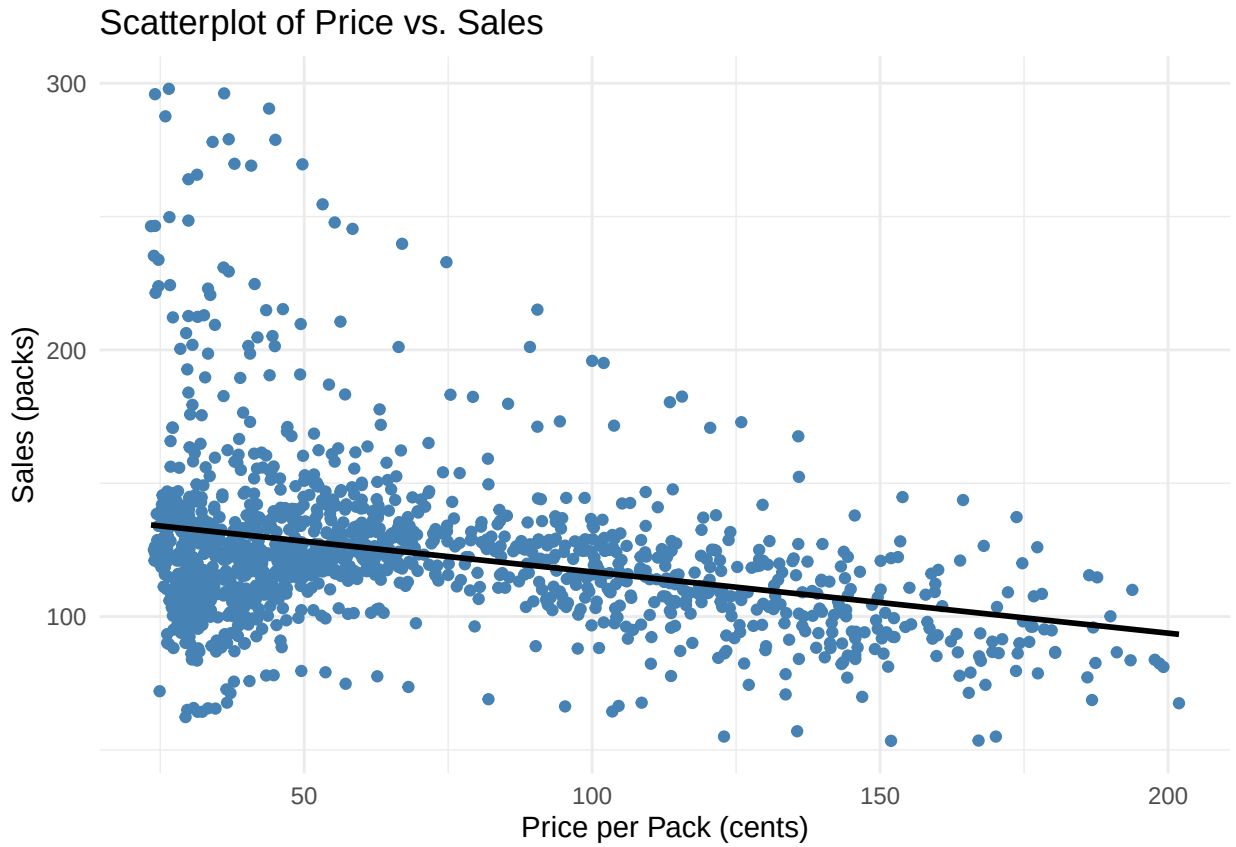
The histogram of per capita cigarette sales reveals a right-skewed distribution, with most observations clustered around the mean but a significant tail at higher consumption levels. This indicates that while the majority of state-year pairs have moderate cigarette consumption, a minority of states or periods experience much higher usage. Such skewness points to the presence of outliers or unique state-level factors, reinforcing the importance of accounting for individual heterogeneity in the analysis. This distributional insight justifies the use of models that can accommodate both typical and extreme consumption behaviors.

2.3 Time Trend of Sales by State



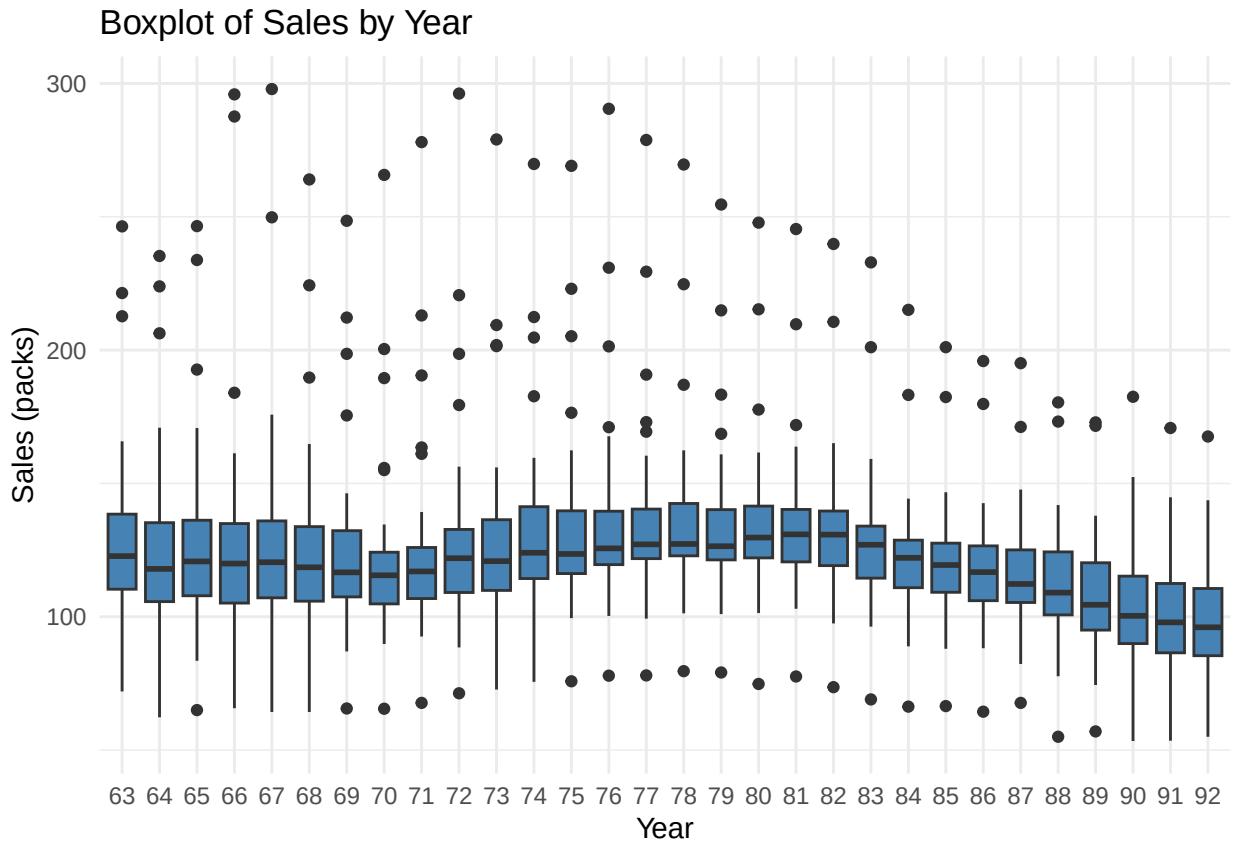
The time trend plots for each state illustrate both a general decline in cigarette sales over the 30-year period and considerable variation in the rate and timing of this decline across states. While the overall pattern is downward, some states exhibit sharper reductions or periods of stability, suggesting the impact of state-specific policies, cultural factors, or economic shocks. These heterogeneous trends highlight the value of panel data techniques, which allow for the separation of common time effects from state-specific influences, ensuring more accurate estimation of the effects of price and income on cigarette consumption.

2.4 Scatterplot: Price vs Sales



The scatterplot depicting the relationship between price and sales shows a clear negative association: as the price per pack increases, per capita cigarette sales decrease. This inverse relationship aligns with economic theory regarding the price elasticity of demand. However, the scatter also reveals considerable spread, indicating that factors beyond price—such as income, demographics, or unobserved state characteristics—may also play significant roles in determining consumption. This underscores the necessity of multivariate panel regression to isolate the specific effect of price while controlling for confounders.

2.5 Boxplot: Sales by Year



The boxplot of sales by year demonstrates a steady decline in median cigarette consumption over time, accompanied by a reduction in the interquartile range. This visual evidence supports the existence of a secular trend toward lower smoking rates, likely driven by increasing health awareness, policy interventions, and shifting social norms. The narrowing dispersion suggests that differences in consumption across states are diminishing, possibly reflecting the diffusion of anti-smoking policies or broader cultural changes. These findings emphasize the importance of including time effects in the regression analysis.

2.6 Correlation Matrix

```
##           sales      price        ndi
## sales  1.000000 -0.3112580 -0.1838390
## price -0.311258  1.0000000  0.9443112
## ndi   -0.183839  0.9443112  1.0000000
```

The correlation matrix indicates that per capita sales are negatively correlated with both price (-0.31) and income (-0.18), while price and income are highly positively correlated (0.94). The negative correlation between sales and price supports the hypothesis that higher prices deter cigarette consumption, while the weaker negative correlation with income suggests a more complex relationship between affluence and smoking behavior. The strong positive correlation between price and income implies that wealthier states tend to have higher cigarette prices, possibly due to higher taxes or cost-of-living. These interrelationships highlight the need for careful multivariate modeling to disentangle the individual effects of each variable on cigarette consumption.

3 Part C: Model Fitting

In this section, we estimate three models to explain the determinants of per capita cigarette consumption across U.S. states from 1963 to 1992. Specifically, we estimate:

- A Pooled Ordinary Least Squares (OLS) model,
- A Fixed Effects (FE) model, and
- A Random Effects (RE) model.

We compare these models using statistical tests, including the F-test, LM test and Hausman test, to identify the most appropriate specification.

3.1 1. Pooled OLS Model

```
##
## t test of coefficients:
##
##                Estimate  Std. Error t value  Pr(>|t|)
## (Intercept) 138.4803259   5.6925561 24.3266 < 2.2e-16 ***
## price        -0.9384115   0.1711029 -5.4845 4.928e-08 ***
## ndi           0.0066364   0.0014924  4.4470 9.410e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
##  studentized Breusch-Pagan test
##
## data:  sales ~ price + ndi
## BP = 27.639, df = 2, p-value = 9.958e-07
```

Conclusion: The Breusch-Pagan test indicates the presence of heteroskedasticity ($p < 0.05$). Thus, we use cluster-robust standard errors, clustered at the state level.

3.2 2. Fixed Effects Model

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = sales ~ price + ndi, data = pdata, effect = "individual",
##     model = "within")
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -74.42972  -6.80022   0.70048   6.79192 126.13990
##
## Coefficients:
##          Estimate  Std. Error  t-value  Pr(>|t|)
## price -0.41288112  0.03753401 -11.0002 < 2.2e-16 ***
## ndi    0.00189843  0.00033904   5.5994  2.61e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    412530
## Residual Sum of Squares: 299900
## R-Squared:      0.27303
## Adj. R-Squared: 0.24737
## F-statistic: 250.125 on 2 and 1332 DF, p-value: < 2.22e-16


##
##  F test for individual effects
##
## data:  sales ~ price + ndi
## F = 73.829, df1 = 45, df2 = 1332, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Conclusion: The F-test rejects the null hypothesis ($p < 0.01$), showing that fixed effects are necessary and preferred over the pooled OLS model. This controls for unobserved state-level heterogeneity.

3.3 3. Random Effects Model

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = sales ~ price + ndi, data = pdata, model = "random")
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Effects:
##                   var std.dev share
## idiosyncratic 225.15   15.00 0.334
## individual    448.39   21.18 0.666
## theta: 0.8717
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -66.40512  -7.36279   0.37018   6.15807 130.90869
##
## Coefficients:
##                Estimate  Std. Error  z-value  Pr(>|z|)
## (Intercept) 138.0289361   3.2359323  42.6551 < 2.2e-16 ***
## price        -0.4273823   0.0375320 -11.3872 < 2.2e-16 ***
## ndi           0.0020310   0.0003389   5.9928 2.063e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    427540
## Residual Sum of Squares: 313430
## R-Squared:      0.26689
## Adj. R-Squared: 0.26582
## Chisq: 501.291 on 2 DF, p-value: < 2.22e-16
```

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  sales ~ price + ndi
## chisq = 8708.5, df = 1, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

Conclusion: The LM test rejects the null ($p < 0.01$), so the Random Effects model is preferred over the Pooled OLS.

3.4 Hasuman Test for Fixed Effects vs Random Effects

```
##
##  Hausman Test
##
## data:  sales ~ price + ndi
## chisq = 183.7, df = 2, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Conclusion: The Hausman test rejects the null hypothesis ($p < 0.05$), showing that the RE assumptions do not hold. Therefore, the Fixed Effects model has better and consistent estimates than the RE model.

3.5 Model Comparison

```
##
## Comparison of Pooled, Fixed, and Random Effects Models
## ================================================================================
##                                    Dependent variable:
##               ---------------------------------------------------------------
##                                          sales
##                      Pooled OLS               Fixed Effects       Random Effects
## --------------------------------------------------------------------------------
## price                 -0.938***                 -0.413***           -0.427***
##                        (0.054)                   (0.038)             (0.038)
##
## ndi                    0.007***                  0.002***            0.002***
##                        (0.0005)                  (0.0003)            (0.0003)
##
## Constant              138.480***                                    138.029***
##                        (1.427)                                      (3.236)
##
## --------------------------------------------------------------------------------
## Observations            1,380                     1,380               1,380
## R2                      0.209                     0.273               0.267
## Adjusted R2             0.208                     0.247               0.266
## F Statistic  181.704*** (df = 2; 1377) 250.125*** (df = 2; 1332)   501.291***
## ================================================================================
## Note:                                                *p<0.1; **p<0.05; ***p<0.01
```

Final Model Selection The Pooled OLS model is rejected due to heteroskedasticity. The Random Effects model is failed from the Hausman test, showing that there is correlation between effects and regressors. The

Fixed Effects model controls for state-level heterogeneity, and is chosen from the F-test, and the Hausman test.

Preferred Model: Fixed Effects Model The Fixed Effects model provides consistent and unbiased estimates for the relationship between cigarette prices, income, and consumption, while controlling for state-level differences over time

4 Part D

4.1 Our Models Findings

When analyzing our preferred model, the Fixed Effects Model, we can determine that both income level and cigarette prices play a notable role in influencing the per capita cigarette consumption across the United States.

4.2 Income Effect

The positive coefficient we get for income (0.0019) suggests that while the effect is small, an increase in income is associated with an increase in cigarette sales. In fact, for a $1,000 increase in income, consumption would rise by 1.9 packs of cigarettes or almost 2 whole packs.

4.3 Price Elasticity

The second and more noticeable influence on cigarette consumption is seen with the price elasticity of cigarettes. The negative coefficient (-0.413) we get for price in our model tells us that a 10% increase in the cost of cigarettes would result in a 4.13% reduction in consumption, which is pretty significant.

4.4 Policy Suggestions

Some policies we could implement to reduce cigarette consumption based on our findings could include taxation and price collaboration between neighboring states. If we implemented a progressive tax increase on cigarettes, particularly in states or counties with higher smoking rates, this could help to reduce consumption per capita throughout each state. Additionally, because states differ in excise tax levels and cigarette prices, getting neighboring states or border counties to implement uniform cost levels would prevent many cross-border purchases in low-tax states.

4.5 Data Limitations

The limitations we must consider when analyzing our results today include the data range we used, along with omitted variables. The data we used in our model spanned from 1963-1992, which represents a much different society than we live in today. More recent data may show significantly different results as vaping has seemed to dominate the nicotine market today. Additionally, we aren't able to consider many variables with our data, such as public health campaigns that have evolved over the years, and shifts in our culture over time.