This lab is an exercise in running fixed effects regression, and specifically in considering functional form. The dependent variables we are going to examine is the airfare of various routes from 1997 to 2000. We will use *fare* as well as *lfare* and *passen* as well as *lpassen.*

Step 1: Consider the variables in the dataset. We know that during the time period these data were collected, a policy was passed intended to reduce anticompetitive practices. In this context, it means that the policy was intended to result in higher prices on hub routes that are more popular. There are 4596 observations in your dataset.

The variables below are the ones in your dataset:

- **year:** 1997, 1998, 1999, 2000
- **id:** route identifier
- **dist:** distance, in miles
- **passen:** avg. passengers per day
- **fare:** avg. one-way fare, $
- **bmktshr:** fraction market, biggest carrier
- **ldist:** log(distance)
- **y98:** =1 if year == 1998
- **y99:** =1 if year == 1999
- **y00:** =1 if year == 2000
- **lfare:** natural log of fare
- **ldistsq:** square of distance
- **concen:** same as bmktshr
- **lpassen:** natural log of avg. passengers per day

Generate summary statistics for your numeric variables for each *year*:

```
library(dplyr)

library(tidyr)


# Step 1: Identify numeric columns except 'year'

numeric_vars <- names(Week11Airfare)[sapply(Week11Airfare, is.numeric) &
names(Week11Airfare) != "year"]


# Step 2: Pivot longer, group by year and variable, then summarize

summary_table <- Week11Airfare %>%
```

```
  select(year, all_of(numeric_vars)) %>%
  pivot_longer(-year, names_to = "variable", values_to = "value") %>%
  group_by(year, variable) %>%
  summarise(
    n = sum(!is.na(value)),
    mean = mean(value, na.rm = TRUE),
    sd = sd(value, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  arrange(year, variable)


# Step 3: Print the result
print(summary_table, n = Inf)  # show all rows
# Nicely formatted console table (optional)
library(knitr)
knitr::kable(summary_table, digits = 2, caption = "Summary Stats by Year and
Variable")
```

Step 2: What did your homework tell you?

In the assignment due for today, you considered what happened if you ignored time and if you included it.

a. What are the major takeaways from the models you conducted?

b. Consider takeaways from both models to write out what you think will happen when we run a formal panel data regression.

Step 3: What did the prework video tell you?

What did the before-and-after regression suggest? What were limitations of the before-and-after method?

Step 4: Run fixed effects models that explore using untransformed and transformed variables.

```
install.packages("fixest")
library(fixest)

# Original model
```

```
m1 <- feols(fare ~ passen + bmktshr | id + year, data = Week11Airfare,
cluster = ~id)

# Log-transformed model

m2 <- feols(lfare ~ lpassen + bmktshr | id + year, data =
Week11Airfare, cluster  = ~id)

# Side-by-side table

etable(m1, m2)
```

a. Interpret each model. First, how does log transforming the dependent variable in the second model change the interpretation? How would you explain what each model tells us? *Hint! You may need to go back to earlier material and check out the table of how to interpret linear, log-log, and log-linear models and coefficients!*

Model 1: Linear and Linear model

Coeffcient passenger Interpretation: Holding everything else constant, for each additional passenger per day on a route, the average fare decreases by $0.0678.This shows a small negative relationship between number of passengers and fare.

Coeffcient bmtshr Interpretation: A 1-unit increase (i.e., 100 percentage point increase) in market share of the biggest carrier increases fare by $7.75. However, this is not statistically significant ($p > 0.05$), so we can't confidently say there's an effect here.

Model 2:
Interpretation Type: Log-linear model

Coefficient on lpassen: -0.3696

Interpretation: A 1% increase in the average number of passengers is associated with a 0.3696% decrease in fare, on average, holding other variables constant.

This is a much more elastic interpretation and statistically significant.

Coefficient on bmktshr: 0.1500

Interpretation: A 1-unit (100%) increase in the market share of the biggest carrier leads to a 15% increase in fare.

This is statistically significant, unlike in Model 1.

b. How does this model seem different than the ones you ran in the homework assignment and the before-and-after regression? Which one do you think is the most useful, and why?

Our homework assignment only focus on R2 for the dependent variable where as this model focuses on the Within R2 too .

c. The output shows you R^2 and Within R^2. These measures tell us different things:

Type of R^2	What it measures	Model 1- interpret in words	Model 2 – interpret in words
Overall R^2	How well the model fits the *entire variation* in the dependent variable	The model explains 95.8% of the total variation in airfare across all routes and years.	The model explains 97.5% of the total variation in airfare across all routes and years.
Within R^2	How well the model explains *within-unit (e.g., within-route)* variation over time	Only 16.9% of the variation in fare over time within the same route is explained by the model. About 43.5% of the variation in log fare over time within the same route is explained.	About 43.5% of the variation in log fare over time within the same route is explained.

You have a high overall R^2 and low within-R^2, meaning that most of the variation in the outcome (e.g., *lfare*) comes from differences across routes, not from changes within a route over time. This means that the fixed effects (route dummies) explain most of the variation. Your predictors (like *lpassen* and bmktshr) explain some of the variation over time within each route, but not much. This could be because the units are stable over time (so, if route characteristics don't change much over the period of time in our dataset). This gap suggests your predictors are better at explaining cross-sectional differences than within-unit changes over time.

d. Can you think of any other reasons that the unseen route dummies would predict more variation than the predictors?

The route dummies are actually the fixed effects . The fixed effects (route dummies) explain more variation than the observed predictors likely because there are important, unobserved, and stable characteristics specific to each route — such as route profitability, type of travelers, local competition, and airport-level infrastructure — that do not vary over time but have a strong impact on fare. Additionally, if predictors like lpassen and bmktshr vary little over time or suffer from measurement error or reverse causality, their ability to explain within-route changes in fare will be limited.

d. Bonus! In the prework video, we looked at scatterplots of fare and passen for 1997, 2000, and before and after. Did log transforming help for the fixed effects regressions? Do you think it would have helped in the before-and-after? Here is some example code to look at a scatterplot; can you prepare code for log variables in 1997, 2000, change and see if they are better? What is the interpretation? What does this tell you about whether before-and-after tells us what we need to know?

```
# Create a scatterplot with lpassen on the x-axis and lfare on the y-
axis
plot(Week11AirfareBA$lpassen2000, Week11AirfareBA$lfare2000,
     xlab = "Log of Passengers 2000",
     ylab = "Log of Fare 2000",
     main = "Scatterplot of Log Fare 2000 vs Log Passengers 2000",
     pch = 16,     # Use filled circles for points
     col = "blue") # Optional: Set color for points
```