# Estimating the Effect of Owning an EV on Annual Miles

Coding Exercise Module 2
*Causal Inference with Linear Regression: A Modern Approach* 2025 by CausAI

## Introduction

At first glance, electric vehicles (EVs) seem like a clear solution for reducing carbon emissions and promoting a greener future. Over the past few years, governments worldwide have encouraged EV adoption through subsidies, tax incentives, and investments in charging infrastructure. The main idea behind these policies is that EVs produce fewer emissions than gasoline-powered cars and are therefore better for the environment.

However, consider a scenario where you work at a government agency studying the broader environmental effects of EV adoption. One concern is that owning an EV might encourage people to drive more. Since EVs generally have lower fueling costs and are perceived as environmentally friendly, drivers may feel less financial or ethical pressure to limit their trips. If EV owners increase their driving, it could offset some of the expected environmental benefits by contributing to road congestion, infrastructure wear-and-tear, and additional emissions from electricity generation.

To examine this issue, you decide to investigate whether individuals who own an EV tend to drive more miles annually compared to those who drive gasoline-powered vehicles.

## Dataset and Variables

To conduct this analysis, you will use a (simulated) dataset collected by the government agency. This dataset contains annual records on driving behavior and characteristics of 89863 American individuals in a given year.

The dataset includes the following variables:

- **Owns EV** (`owns_ev`): A binary variable indicating whether an individual owns an electric vehicle (1) or not (0) in a given year. (assumed to be fixed throughout the year: either someone owned an EV the entire year, or not at all)
- **Miles Driven** (`miles_driven`): The total number of miles an individual has driven over the year.
- **Gas Price** (`gas_price`): The price of gasoline during the corresponding year, assumed to remain constant throughout that year.
- **Fixed Expenses** (`fixed_expenses`): The total amount spent on fixed costs such as rent, insurance, and subscriptions in U.S. dollars for the corresponding year.
- **Age** (`age`): The individual's age in years.
- **Previous Year's Income** (`last_year_income`): The individual's annual income in the preceding year, also in U.S. dollars.
- **Gender** (`gender`): A binary variable indicating the individual's gender, where 0 represents male and 1 represents female.
- **Urban vs. Rural Living** (`urban_rural`): A binary variable indicating whether the individual resides in an urban (0) or rural (1) area. This status is assumed to be unchanged for the previous and this year.

# Objective

In the upcoming modules of this course, the goal will be to estimate the *average treatment effect of owning an electric vehicle (EV) on annual miles driven.* Specifically, we aim to determine whether EV ownership leads to an increase in the average number of miles driven per year. Understanding this relationship is important, as it can inform policy decisions related to EV adoption. If owning an EV instead of a gasoline car increases miles driven, this should be taken into consideration for future policy recommendations to avoid that we overestimate the environmental benefits of promoting EV ownership.

# Exercises

In the following exercises, you will work with the dataset `"EV_Miles.csv"` to explore the relationship between electric vehicle (EV) ownership and annual miles driven using linear regression. If you are unfamiliar with any of the techniques mentioned (such as z-score normalization or cross-validation), you can ask ChatGPT for help.

1. Download the dataset `"EV_Miles.csv"` and load it into Python.

2. Perform an exploratory data analysis to understand the dataset. This can include computing summary statistics (e.g., mean, median, standard deviation), creating visualizations such as histograms, scatter plots, or box plots, and so on. Basically, anything that will help you get familiar with the data.

3. Calculate the average miles driven separately for EV owners and non-EV owners. Then compute the difference between the two averages. How would you interpret this difference? Do you think this difference represents the average treatment effect of EV ownership on miles driven? If not, what kind of variables can you think of that might exist that produce bias in this associational difference?

4. Run a few linear regressions of `miles_driven` on `owns_ev` and other variables you suspect to be most relevant. Examine the coefficient for `owns_ev` in each model. How would you interpret it?

   (*Optional*: You can standardize (z-score normalize) the continuous variables (excluding the outcome variable) for better numerical stability of the regression. This can easily be done using `StandardScaler` from scikit-learn.)

5. Use 5-fold cross-validation to select the model with the best predictive accuracy, using `owns_ev` and 4 other control variables. More specifically,

   - Generate all possible combinations of 4 variables from the following list (excluding `owns_ev` and `miles_driven`): `gas_price`, `fixed_expenses`, `age`, `last_year_income`, `gender`, and `urban_rural`. You can use Python's `itertools.combinations` for this.
   - For each combination, perform 5-fold cross-validation:
     - Split the data into 5 equal parts (folds)
     - For each fold: train a linear regression model on 4 folds and test on the remaining fold using the Mean Squared Error (MSE) loss.
     - Record the MSE on each test fold, and compute the average MSE over the 5 folds
   - Identify the variable combination that gives the lowest average MSE. This is the best-performing model in terms of predictive accuracy.

   Which 4 variables are included in the best model? What is the coefficient for `owns_ev` in this model, and how would you interpret it?

6. Repeat the previous step, but now using 5-variable combinations. What is the best-performing combination in this case? What coefficient do you get for `owns_ev`? How do you interpret it?

7. Based on your analysis so far, can you draw any causal conclusions about the causal effect of owning an EV on annual miles driven? Why or why not?

Hmm... we still don't know the causal effect of owning an EV on miles driven.
To be continued...