ICCT Final Interview

Colton Lapp
February 16, 2023

Overview

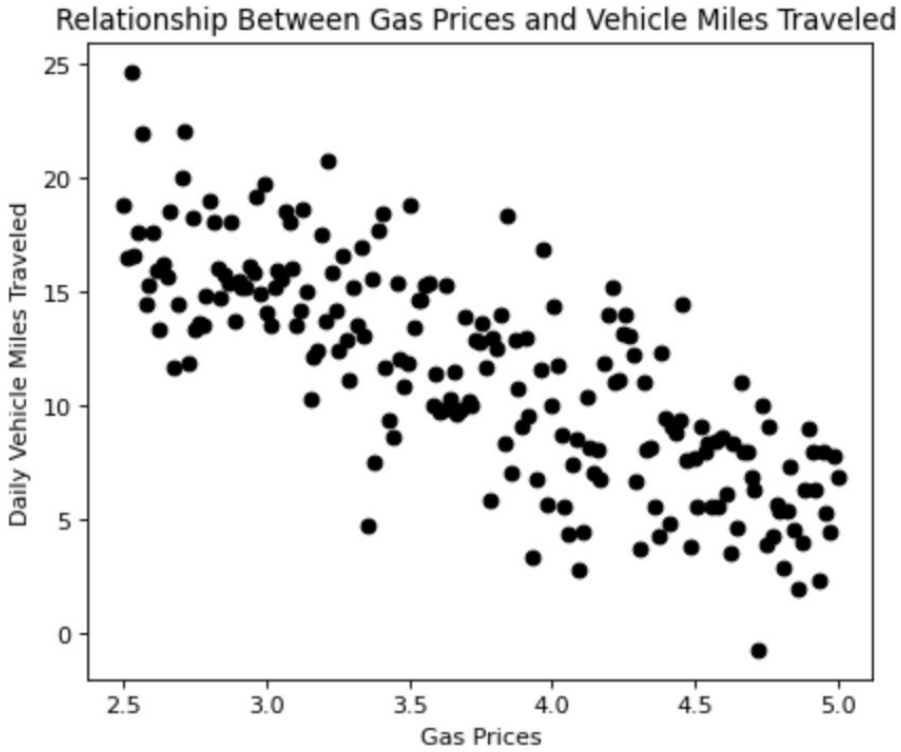
1. Explanation of statistical method: **linear regression**
2. Deployment demo: **how to implement the model**

Goals of talk:

- Demonstrate:
  - Technical communication abilities
  - Knowledge of Python syntax and statistics

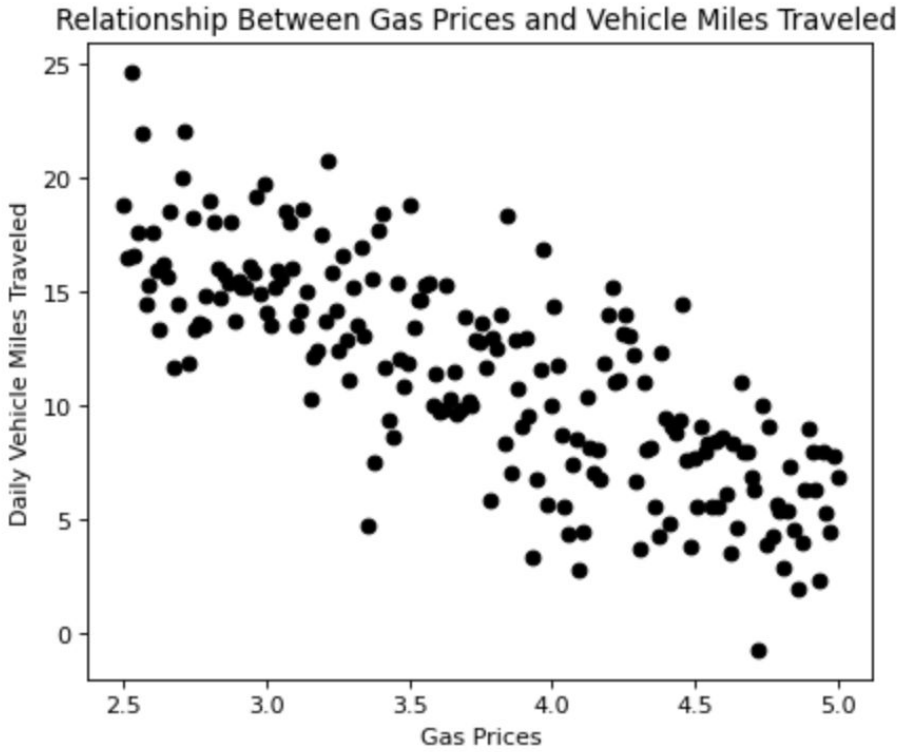
Python script to reproduce figures available on my Github here

Modeling the relationship between two variables



Q: How can we model the relationship between these variables?

The simple linear model

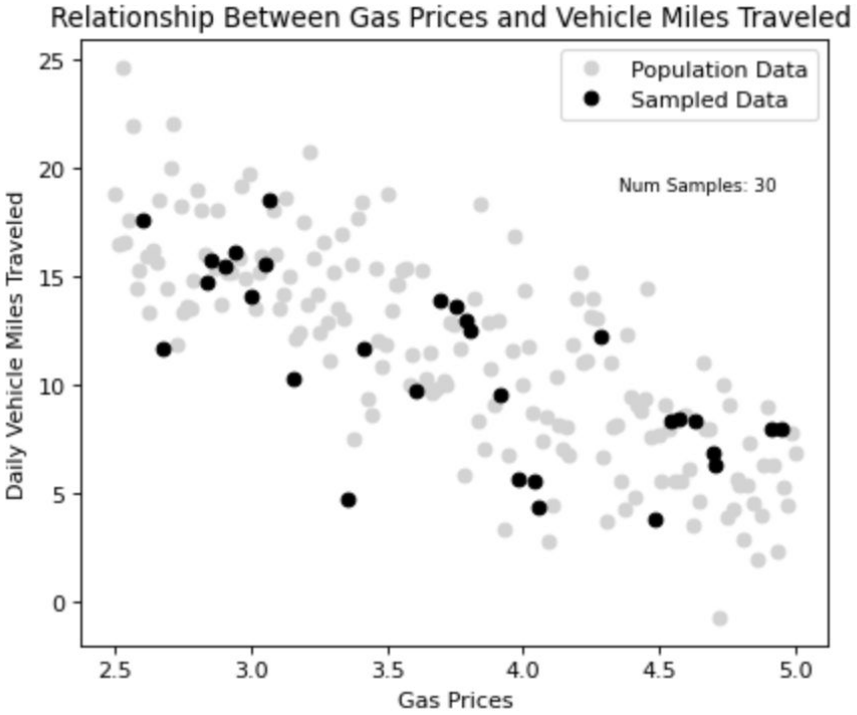


In this example, the population data is simulated from a linear model:

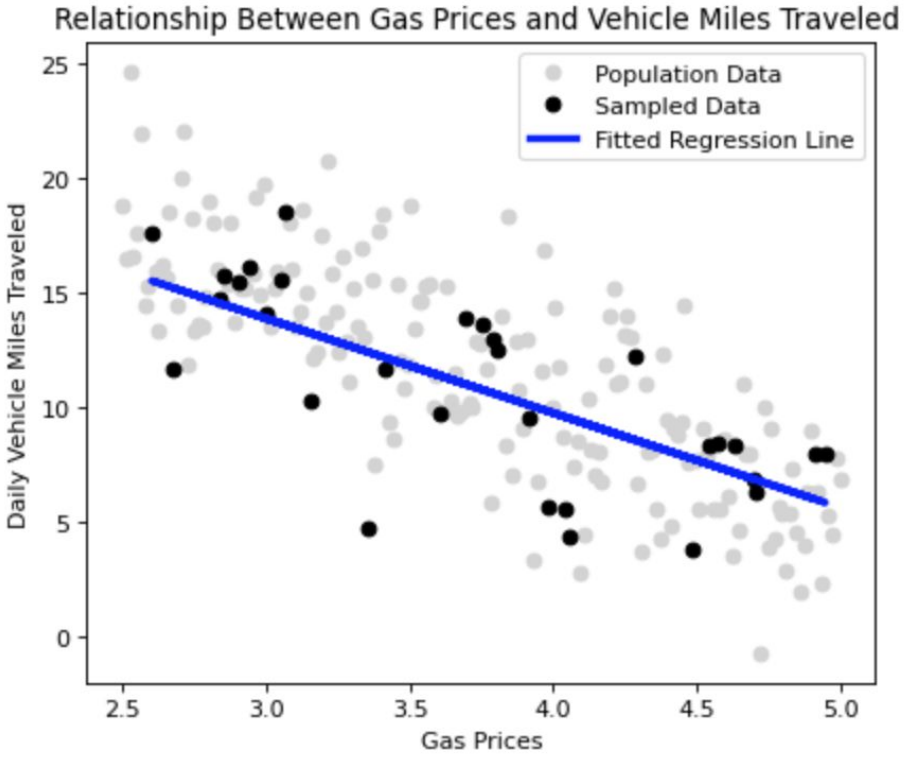
$$Y = \beta_0 + \beta_1 X + u$$

$$\beta_0 = 30, \beta_1 = -5$$

Data is “sampled” from a “population”

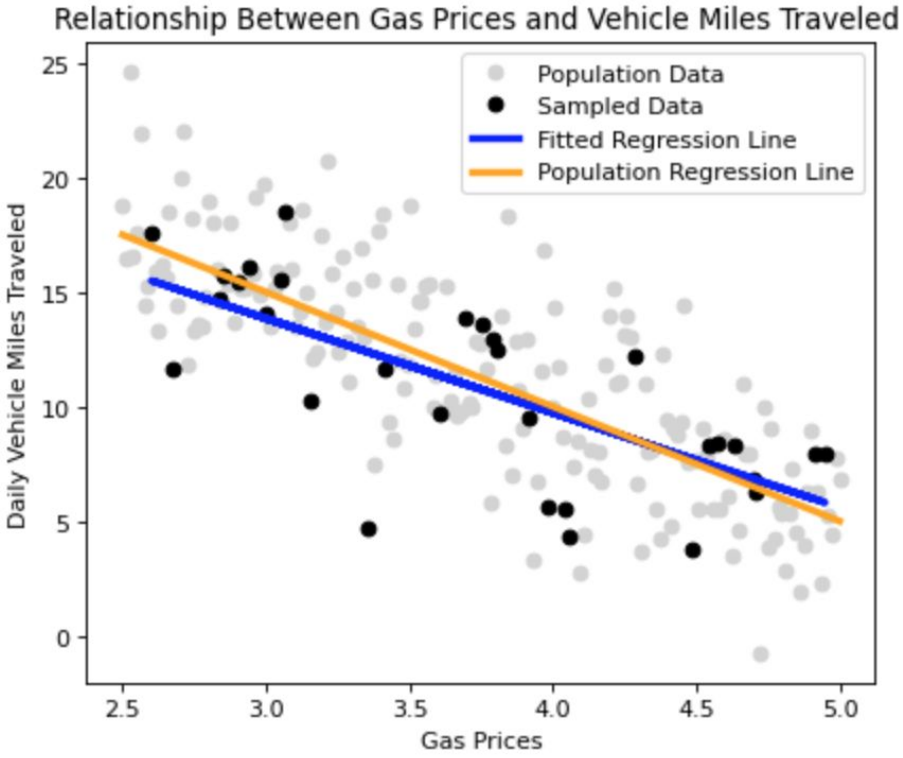


Goal: Estimate coefficients for model



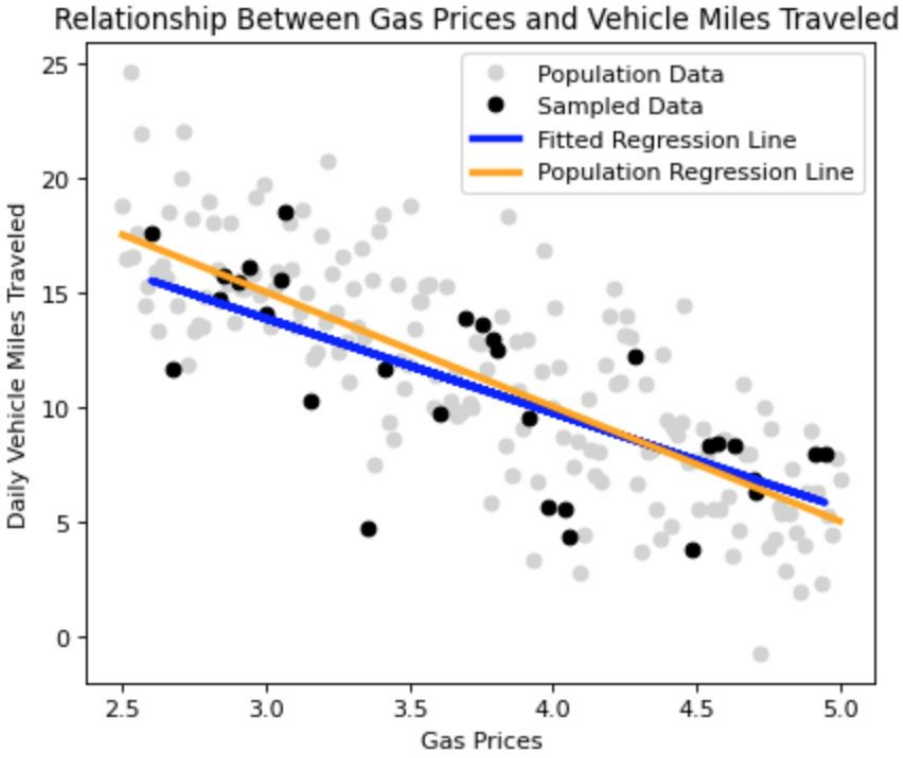
$$Y = \beta_0 + \beta_1 X + u$$

Goal: Estimate coefficients for model



$$Y = \beta_0 + \beta_1 X + u$$

Goal: Estimate coefficients for model

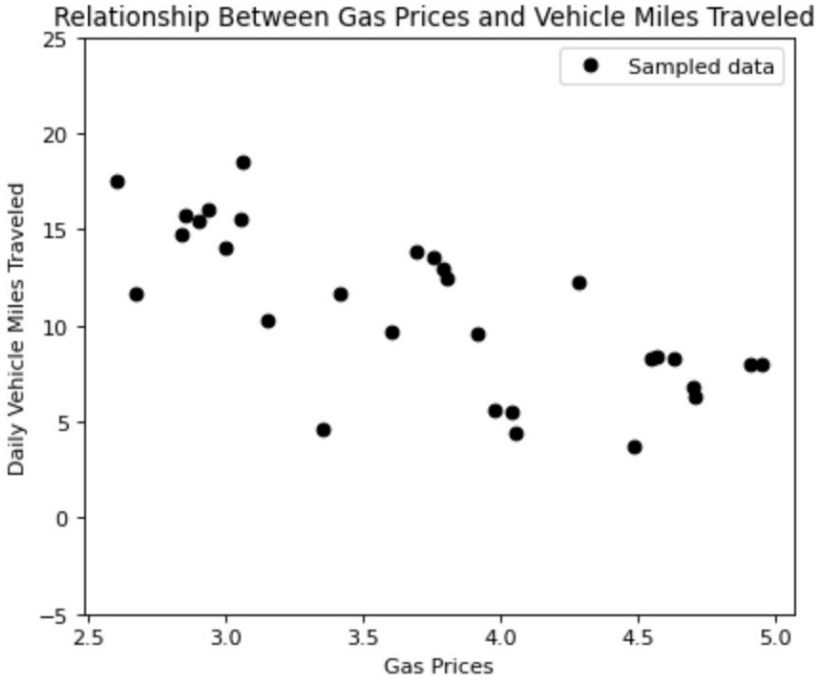


$$Y = \beta_0 + \beta_1 X + u$$

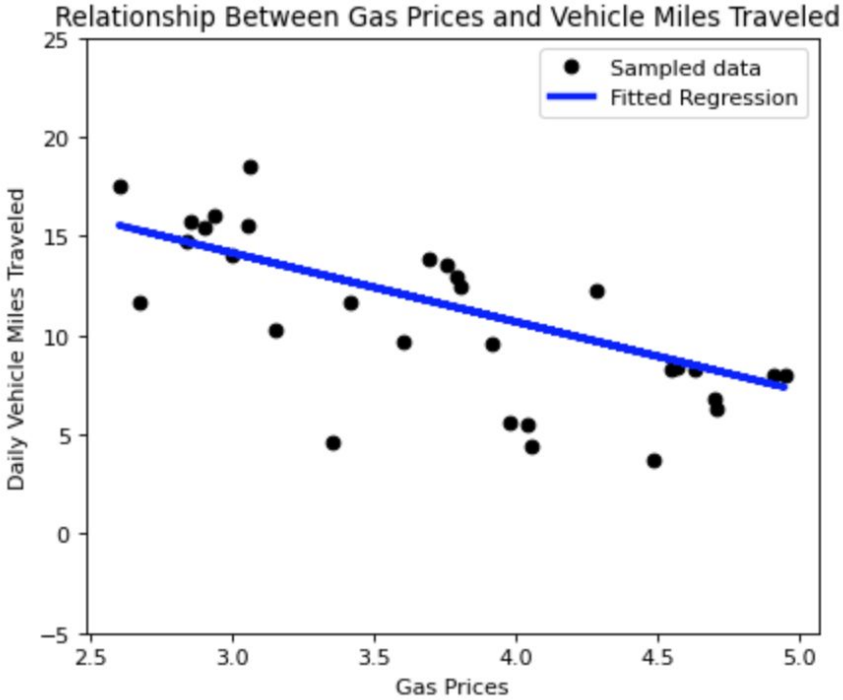
Can accomplish two things:

- Prediction
- Causal Inference

How do we estimate the coefficients?

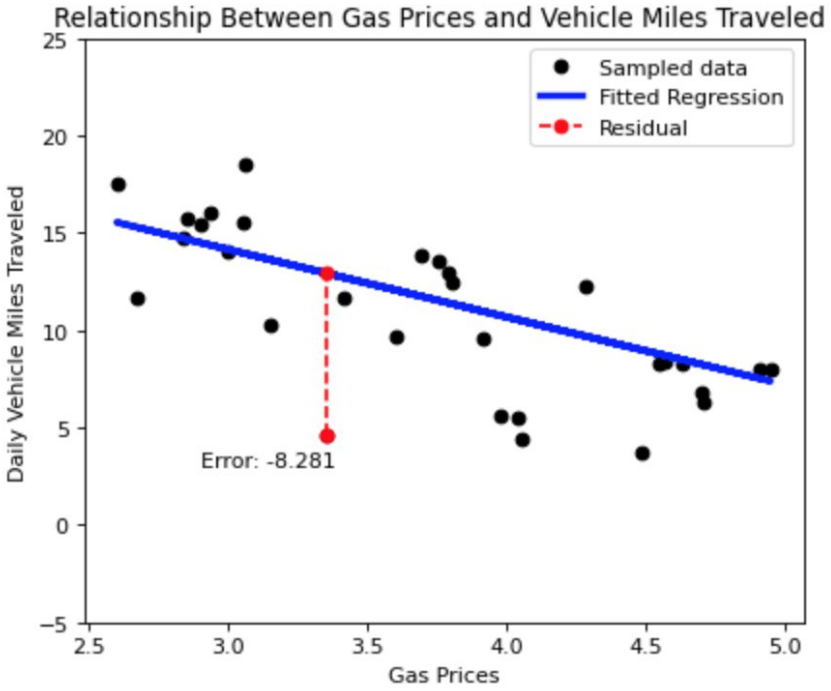


How do we estimate the coefficients?



$$Y = \beta_0 + \beta_1 X + u$$

How do we estimate the coefficients?

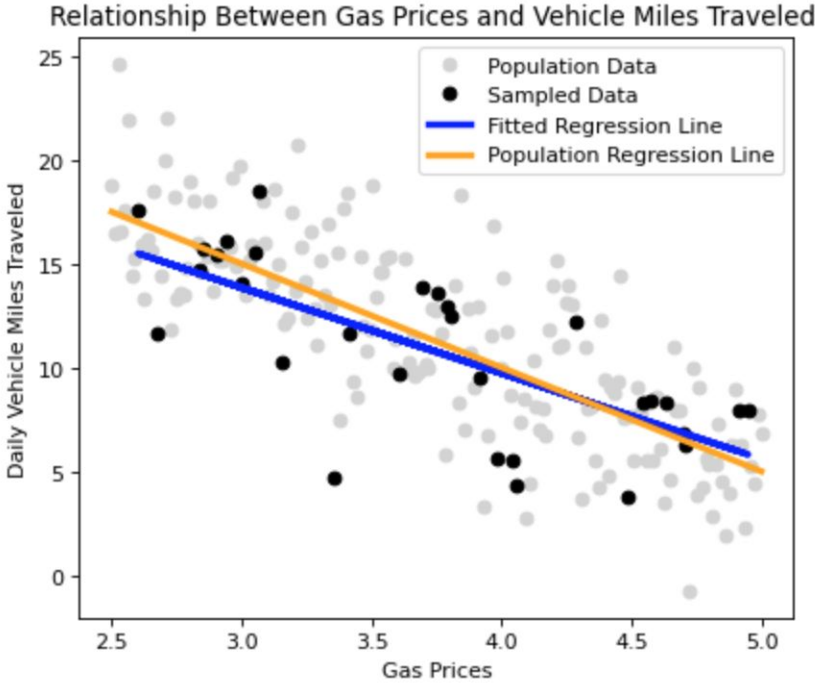


Preferred method: Minimize the “Sum of Squared Errors”

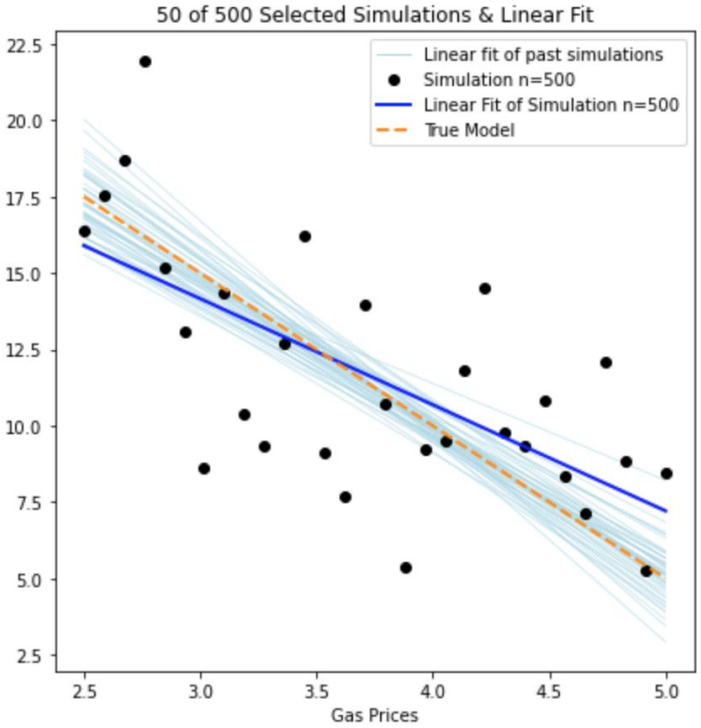
$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{b_0, b_1} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- Can achieve this using:
  - Analytical solution (calculus)
  - Optimization algorithms (i.e. gradient descent)

How accurate are the parameter estimates compared to the “true” model?



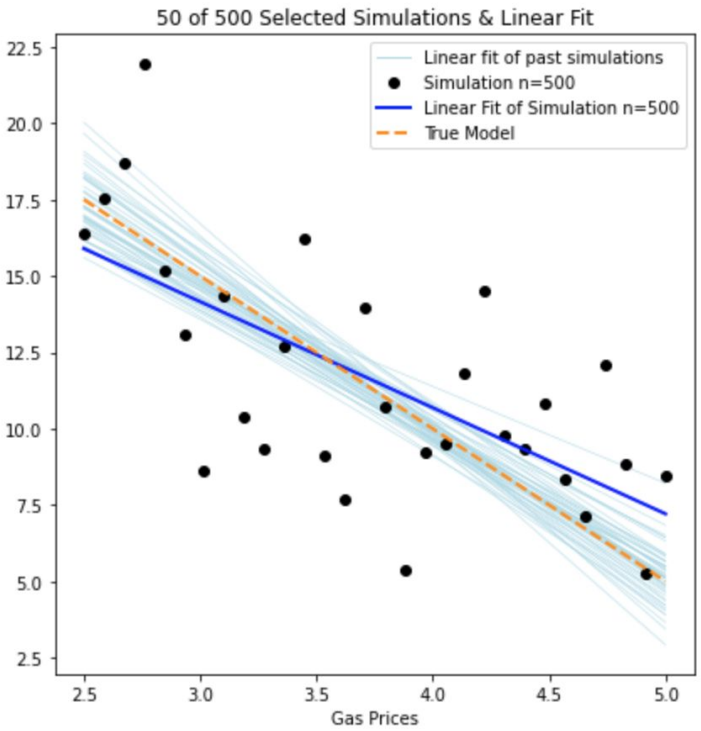
How accurate are the parameter estimates compared to the “true” model?

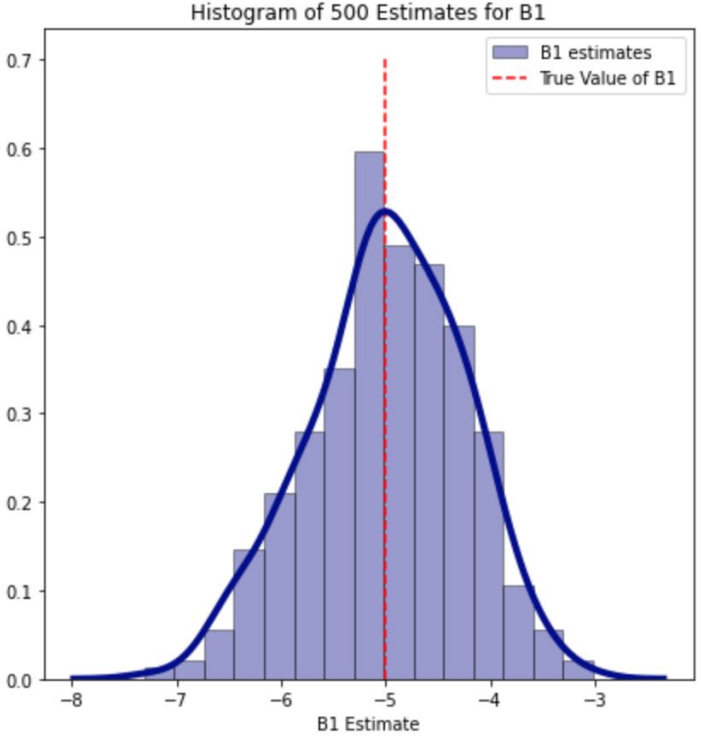


Pseudo Code:

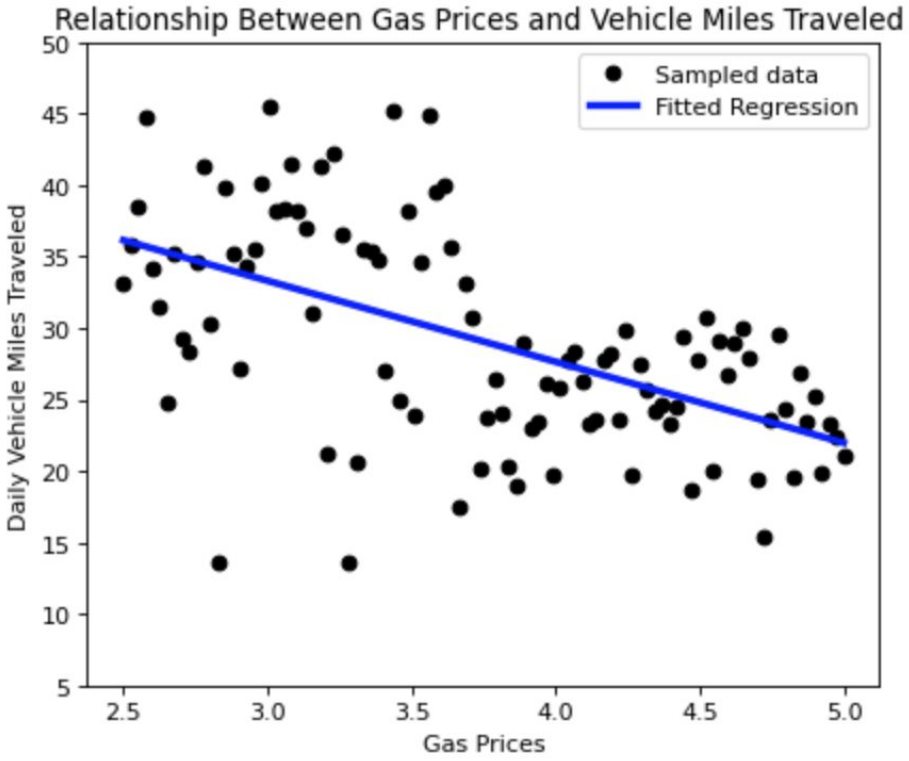
```
For sim in range(0,500):
    -Sample 30 observations from population
    -Estimate parameters
    -Save parameter estimates
```

Under Gauss-Markov assumptions, parameter estimates are statistically well defined





Potential Problems: Omitted Variables



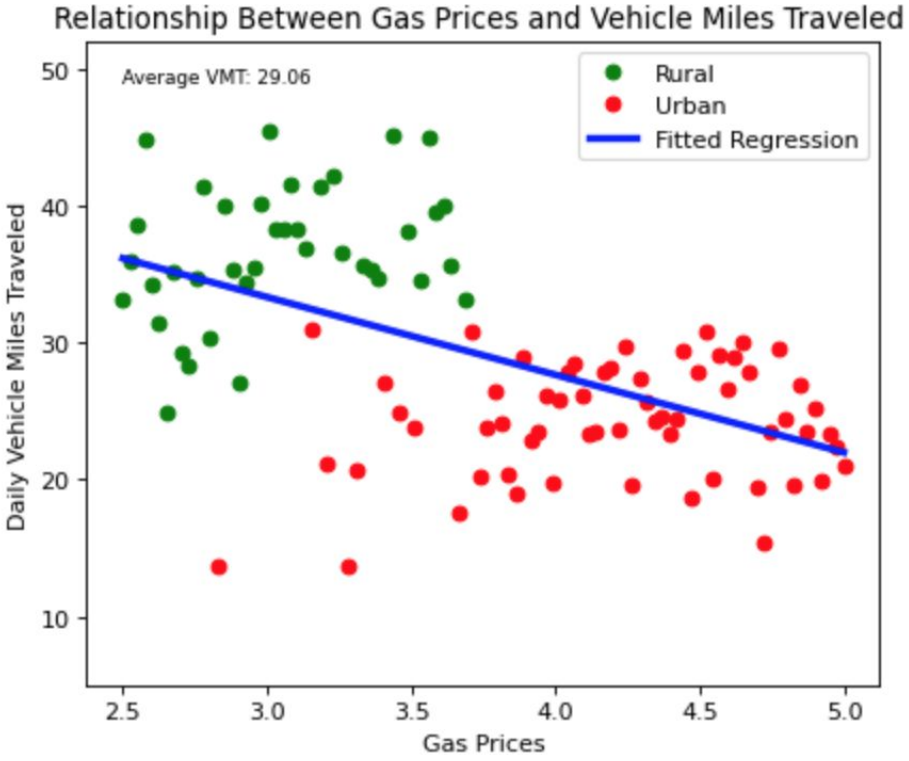
$$Y = \beta_0 + \beta_1 X + u$$

One might conclude that high gas prices "cause" people to drive less

Estimated model:

VMT = 55.3 **- 7.1*Prices**

Potential Problems: Omitted Variables



In reality…

Rural areas have lower gas prices
- and -
Rural areas have higher amounts of VMT

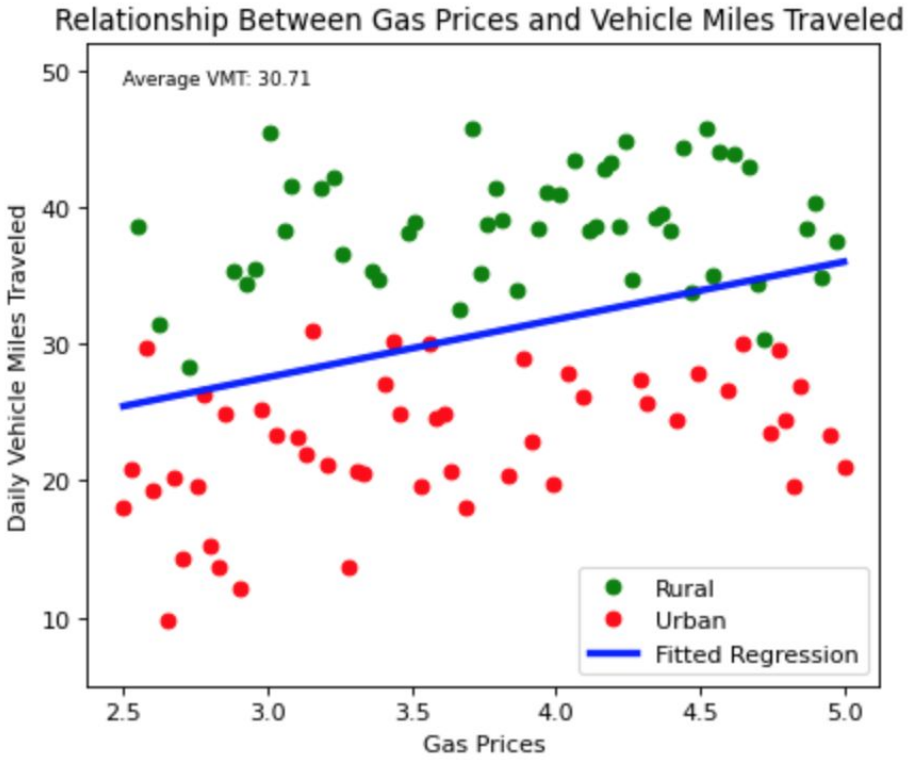
Estimated model:

VMT = 55.3 **- 7.1*Prices**

True model:

VMT = 15 + **15* Rural + 2***Prices

Potential Problems: Omitted Variables



If omitted variables aren't correlated with explanatory variables, however, this isn't an issue!

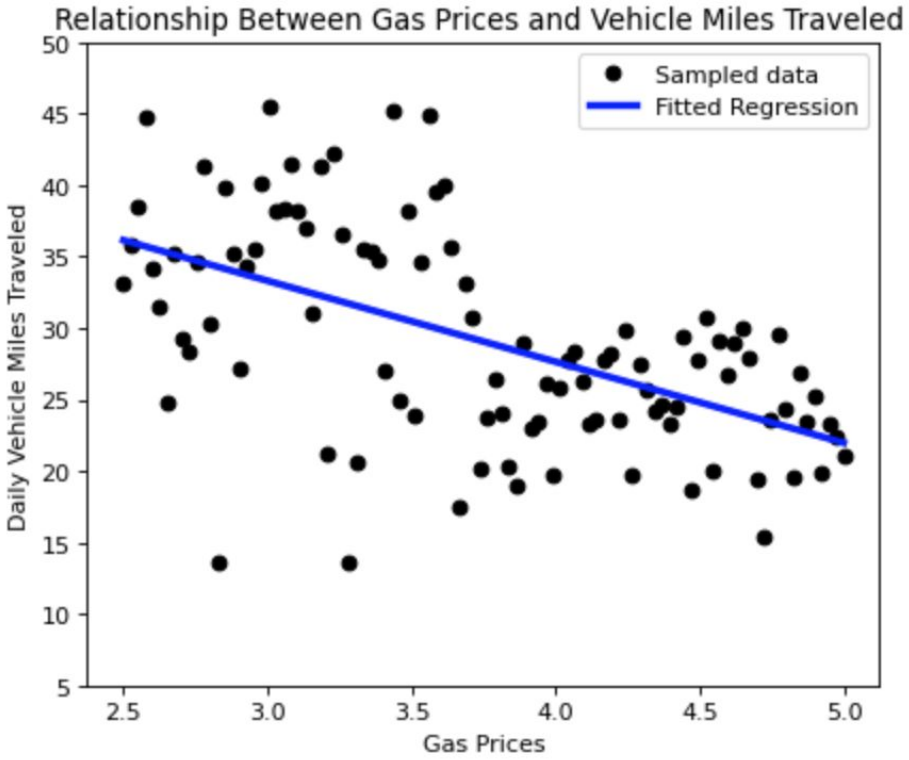
True model:

VMT = 15 + 15* Rural + **2*Prices**

Estimated model:

VMT = 21.4 + **2.5*Prices**

Potential Problems: Omitted Variables



If our goal is only prediction, however, gas prices are still a good proxy for VMT

Brainstorming a Real World Application

Outline:

Goal: Predict reduction in VMT in response to fuel tax

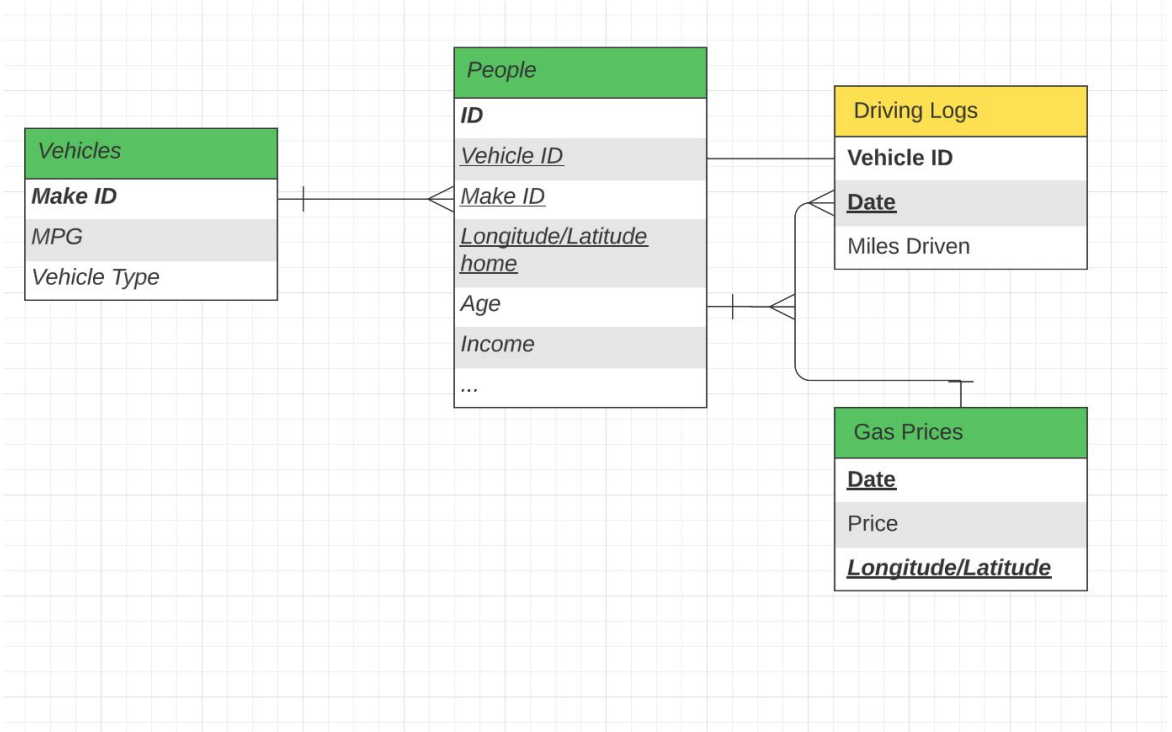
Outline:

Goal: Predict reduction in VMT in response to fuel tax

- **Data**:
  - Relational Database of daily driving logs
- **Regression model:**
  - Linear regression with control variables
- **Python setup:**
  - Creating custom classes to streamline modeling

Database: ERD schematic

Assume we have daily driving logs on a individual level

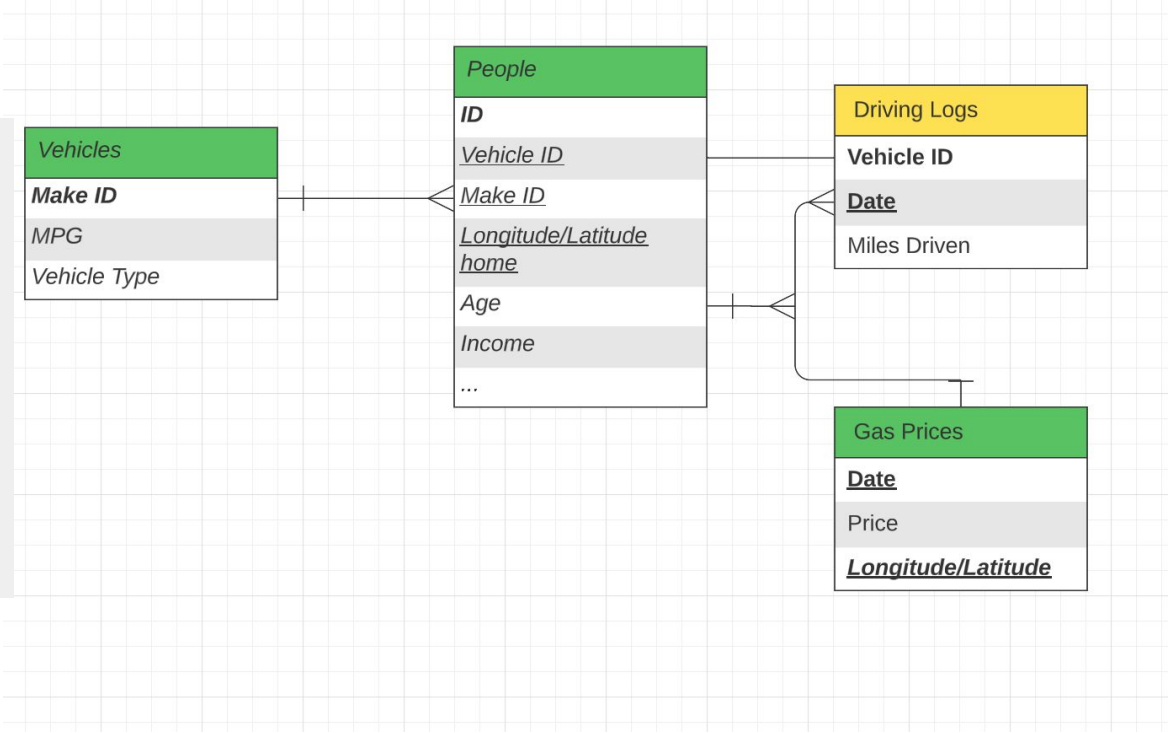


Database: ERD schematic

Assume we have daily driving logs on a individual level

Select Data from DC between 2015-2020:

```
SELECT *
  FROM people p
    JOIN driving_logs dl USING (Vehicle_ID)
    JOIN vehicles USING(Make_ID)
    JOIN gas_prices gp
          ON p.MSA_code = gp.MSA_code
        AND dl.date = gp.date
WHERE dl.date
          BETWEEN '2015-01-01'
          AND '2020-01-01'
    AND  pl.MSA_code = 1;
```



Linear Regression Model: adding controls

Options:

- Specific models for
  - Location
  - Vehicle Type
  - Etc
- Or:
  - Only use control variables

$$VMT_{i,t} = \beta_0 + \beta_1 \text{Price}_{i,t} + \sum_{p=2}^{p'} \beta_p X_p + u_{i,t}$$

$$i \in \{city_1, city_2, \ldots, city_i\}$$

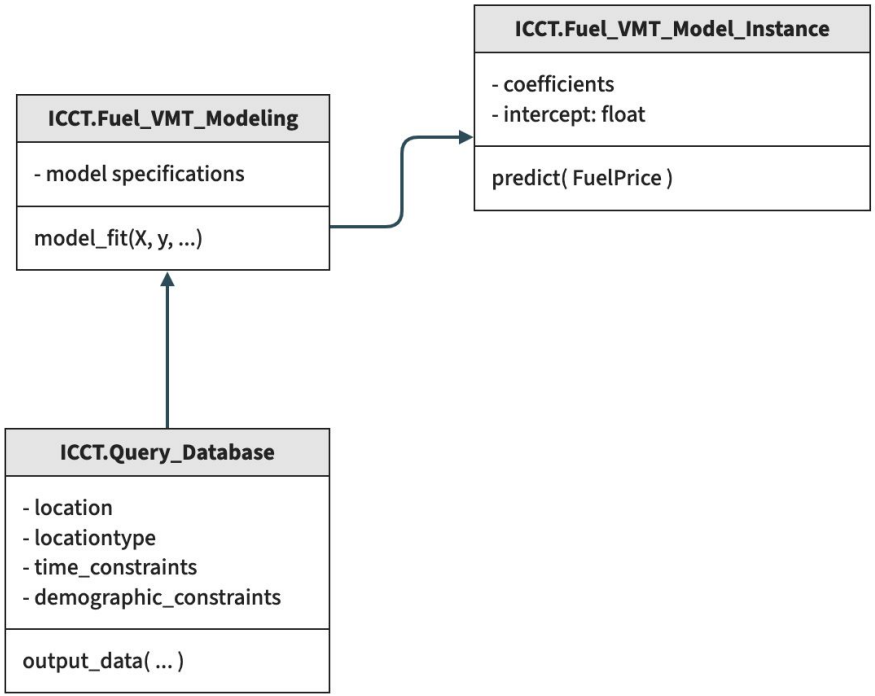
$$t \in \{day_1, day_2, \ldots, day_t\}$$

X_p could be: Demographic data, Vehicle Data, Date data, etc...

Python implementation: UML notation

Could use Python classes to:

- Query database
- Create model
- Use model to predict changes in VMT after changes in fuel prices



Questions?

References

1. https://scholar.princeton.edu/sites/default/files/bstewart/files/lecture5slides.pdf
2. https://jakevdp.github.io/PythonDataScienceHandbook/05.06-linear-regression.html
3. https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf