# STAT 429 Final Project

Samuel Wu (samuel54), Molin Yang (moliny2), Rongsheng Zhang(rz36)

12-10-2025

## Abstract

This project examines how used vehicles prices in the US are related to several key factors: new vehicle CPI, fuel prices CPI, federal funds. Using monthly data imported from federal reserve bank of St. Louis covering Janurary 2000 through August 2025, we model CPI for new cars with the four key factors which we mentioned above. We constructed two linear regression models: a baseline specification and and extended version that incorporates a structural dummy variable (COVID-19). Our results show that new vehicle prices are the strongest predictor of used-vehicle prices, meanwhile fuel costs and federal funds play insignificant roles. In the second model, the COVID-19 dummy variable is highly significant and indicates that the used vehicle prices were roughly 10 index points higher on average after 2020, even after accounting for other predictors. After this, we used the second model to predict the future five month values for used vehicles CPI, which showed a modest continued rise in used vehicles CPI in the near term.

## Introduction

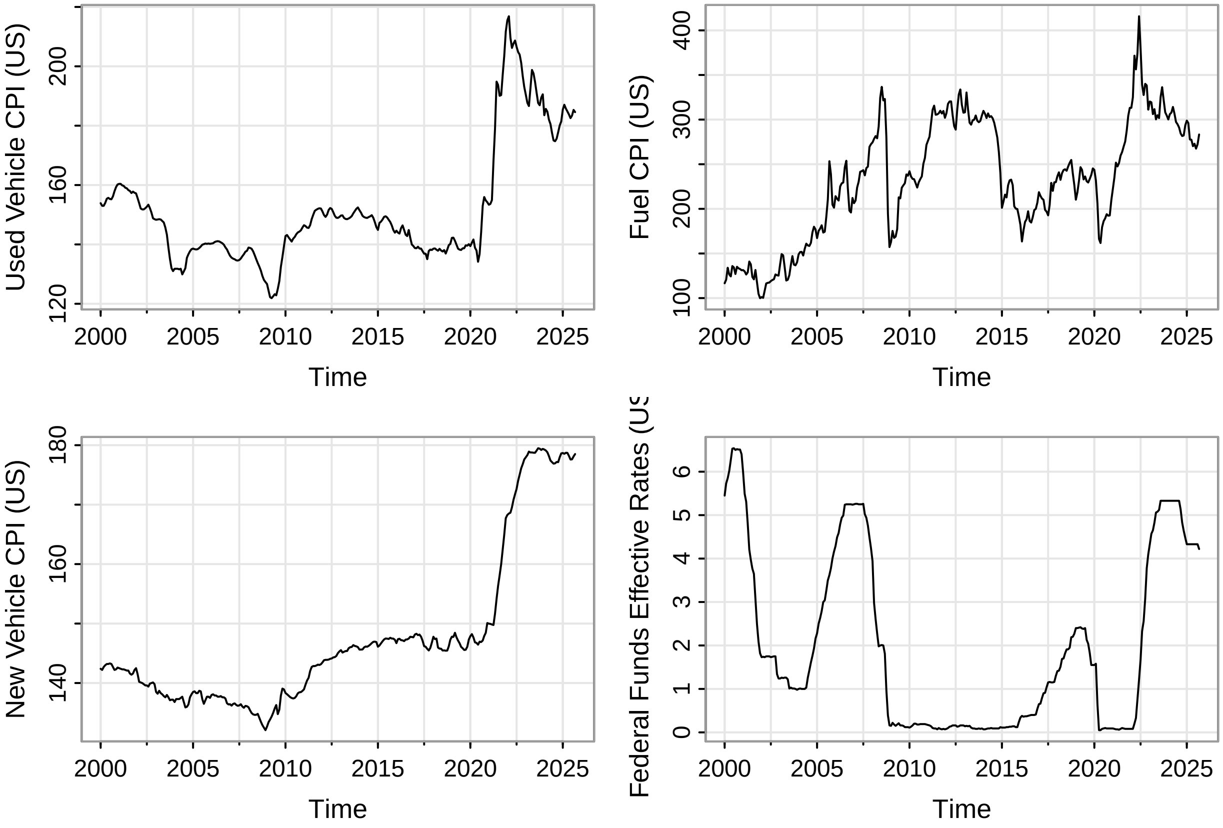
In recent years, movements in consumer prices have become a central concern for households, firms, and policymakers in the United States. Among the various components of the Consumer Price Index (CPI), the index for used cars and trucks is of particular interest. Used vehicles are a relatively inexpensive way for many households to acquire transportation, and huge swings in their prices can immediately affect the cost of commuting and access to employment, as well as broader household budgets. Following the COVID-19 pandemic, the prices of used vehicles rose unusually rapidly and then partly reversed course. These fluctuations were cited extensively as one of the more visible contributors to the recent inflationary episode. This places great importance on understanding how the prices of used vehicles change over time and relate to overall macroeconomic conditions.

The goal of this project is to understand the relationships between used-vehicle prices and several key economic indicators that are plausibly related to the auto market.

In our project, we ask to what extent variation in the CPI for used cars and trucks can be explained by (i) the CPI for new vehicles, which reflects conditions in the primary auto market; (ii) the CPI for motor fuel, which captures the cost of operating vehicles; and (iii) the federal funds effective rate, which summarizes the stance of monetary policy and the general level of short-term interest rates.

Time series visual inspection of the data suggests that these series move together at least part of the time, but it is not obvious from the plots alone which predictors are most influential or how strong the relationships are once we control for common trends. Also we can not find the lag relationships clearly. Statistical time series models are therefore useful in this context: they allow us to separate the effects of different predictors, account for autocorrelation and long-run trends, and provide a model for evaluating how well these variables collectively explain movements in used-vehicle prices. For this project we also want to forecast 5 values in the future based on our model.

Our analysis is based on four monthly U.S. macroeconomic series sourced from the Federal Reserve Bank of St. Louis (FRED) database. The response variable is the CPI for used cars and trucks. The three explanatory variables are the CPI for new vehicles, the CPI for motor fuel, and the federal funds effective rate. All four series are aligned at a monthly frequency starting in January 2000 and continuing through August 2025. This unified time span enables us to study the long - term mutual influences across multiple economic cycles. Meanwhile, it also allows us to analyze the impacts brought about by the rapid changes emerging due to the COVID-19 pandemic.



To provide an initial overview, Figure 1 displays time plots of the four series on a common calendar from 2000 onward. The used-vehicle CPI shows a gradual upward trend with relatively modest fluctuations in the early 2000s, followed by much sharper increases and decreases around the pandemic period. The new-vehicle CPI also trends upward over time but exhibits smoother movements. In contrast, the motor fuel CPI is highly volatile, with repeated spikes and collapses that reflect swings in energy prices. The federal funds effective rate remains low for much of the period following the 2008 financial crisis, drops back near zero at the onset of COVID-19, and then rises rapidly as monetary policy tightens. These patterns suggest that all four series are dynamic, potentially non-stationary, and likely interconnected.

In the main part of the report, we will use time series models to investigate how used-vehicle prices respond to changes in new-vehicle prices, fuel costs, and interest rates. Given the disruptive impact of the COVID-19 shock, we will specifically incorporate January 2020 as a structural breakpoint to compare these economic relationships in the pre-pandemic versus post-pandemic periods. By fitting and comparing alternative model specifications, we aim to assess the relative importance of each predictor, evaluate how well these variables explain historical variation in the used-vehicle CPI, and explore the extent to which the resulting models can be used for short-term forecasting.

# Methods

## Data and Preprocessing

In our project, four monthly time-series datasets from FRED are being used:

- CPI for used cars & trucks (response variable)
- CPI for new vehicles
- CPI for motor fuel
- Federal funds effective date

And here are the reference IDs for each datasets being used in this project:

- series ID CUSR0000SETA02
- series ID CUSR0000SETA01
- series ID CUSR0000SETB
- series ID FEDFUNDS

In order to provide a consistent analysis and matching data points within the datasets, the datasets are being filtered. We cleaned up the datasets which means only data points from January 2000 - August 2025 are being used in this analysis and modeling. After doing so, we noticed that the column for date in the datasets are in string format. This induces us to use R code to convert this column from string format to date format, which formalizes the data and enables further analysis and graphing in the later stage. Lastly, we also renamed the columns in each dataset that consist of the price information appropriately so that later the analysis will be easier.

After doing so to each individual dataset, each data is in consistent format; however, all the datasets are still separated from each other, lacking unison. In order to fit a simple linear regression model, we need to join four datasets together on a primary key, which is the date column. Therefore, we used date as the primary key column and we used the function `left_join` in R to join four datasets together. Hence, we now have a clean dataset consisting of five columns, “date, used_cpi, new_cpi, fuel_cpi, fedfunds”. Following that, we need to construct another two columns for further analysis, “index” and a dummy variable “covid”, 0 for “not in covid time”, 1 for “in covid time”. After doing so, our setup is considered as complete and ready for the next step.

## Model Specification

In this project, our group is going to fit two models. The first model will be a Baseline Multiple Linear Regression. It is written as:

$$\text{used\_cpi} = \beta_0 + \beta_1 \cdot \text{new\_cpi} + \beta_2 \cdot \text{fuel\_cpi} + \beta_3 \cdot \text{fedfunds} + \epsilon$$

The purpose of this model is for us to assess how macro variables jointly explain used vehicle CPI.

The second model which we will construct is a regression but with structural break. This model will be same as Model 1 but adds Covid variable (dummy variable).

The model can be listed as:

$$\text{used\_cpi} = \beta_0 + \beta_1 \cdot \text{new\_cpi} + \beta_2 \cdot \text{fuel\_cpi} + \beta_3 \cdot \text{fedfunds} + \beta_4 \cdot \text{covid} + \epsilon$$

The purpose of fitting this model is to capture a level shift caused by COVID-19 (Since 2020 January). Please note that both models are estimated via ordinary least squares (OLS).

### Model Evaludation and Comparison

After fitting the models, it is important to analyze the results of the model and then compare them in order to evaluate which model is more efficient. In our project, we are going to use the following criteria to evaluate which model are better:

- Adjusted $\mathcal{R}^2$ for goodness of fit.
- AIC for penalized fit
- Statistical significance of individual predictors, p-values.

Besides this, we are going to check residuals visually for general patterns to ensure that the residuals follow a normal distribution.

### Forecasting procedure

We are going to forecast the future 5 values by the following procedure. Firstly, we are going to fit simple time-trend regressions separately for each predictor:

- `new_cpi ~ t`
- `fuel_cpi ~ t`
- `fedfunds ~ t`

After that, we are going to forecast the next five months of each predictor. After forecasting the future five values we are going to save them in a new dataset along with the covid variable = 1. Lastly, we are going to plug values into the preferred model (Model 2) to generate the future 5 values for used CPI.

---

## Results

### Overview of Model Fits

After fitting the model and producing the results, we discovered that both models fit the data reasonably well. Model 1 in this case produced the adjusted R square of 0.7466, which means that the model explains about 75% percent of the variation. After adding the dummy variable Covid, the model is improved. Within all the predictor variables, New vehicles CPI showed strong, consistent significance; however, Fuel CPI and FedFunds show weak significance.

## Model 1 Results.

```
##
## Call:
## lm(formula = used_cpi ~ new_cpi + fuel_cpi + fedfunds, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.017  -7.342  -0.892   4.408  37.321
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -42.241203   6.965645  -6.064 3.92e-09 ***
## new_cpi       1.309713   0.056201  23.304  < 2e-16 ***
## fuel_cpi      0.003378   0.010652   0.317    0.751
## fedfunds      0.029575   0.312600   0.095    0.925
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 10.13 on 305 degrees of freedom
## Multiple R-squared:  0.7491, Adjusted R-squared:  0.7466
## F-statistic: 303.5 on 3 and 305 DF,  p-value: < 2.2e-16
```

Here are the results of the baseline model that our group fit. First and foremost, the coefficient significance presented the following numbers.

- new_cpi -> $3.92 \times 10^{-9}$, way smaller than 0.05.
- fuel_cpi -> 0.751, not significant.
- fedfunds -> 0.925, not significant.

Therefore, we know that new_cpi is the only significant predictor. This coefficient can be explained as every single one point increase in new vehicle CPI is associated with +1.31 increase in used vehicle CPI. Besides this predictor, the other two predictors (Fuel and Fedfunds) have negligible short-run effects in this model.

After that we can analyze the result for the goodness of fit. From the results,

- Adjusted $R^2 = 0.7466$
- Residual SE $\approx 10.13$

This means that the model explains about 75% of the variance. The residual standard error for Model 1 is about 10.1 index points, indicating the typical deviation between the fitted values and the actual used_vehicle CPI.

In summary, New Vehicle CPI is the dominant predictor, and baseline model is decent but misses the structural changes.

## Model 2 Results

```
##
## Call:
## lm(formula = used_cpi ~ new_cpi + fuel_cpi + fedfunds + covid,
##     data = dat)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -23.588  -7.084  -0.544   4.629  35.432 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -3.923299  11.764878  -0.333    0.739    
## new_cpi      1.026557   0.089686  11.446  < 2e-16 ***
## fuel_cpi     0.007246   0.010446   0.694    0.488    
## fedfunds     0.191754   0.307910   0.623    0.534    
## covid       10.014526   2.508951   3.992 8.24e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## 
## Residual standard error: 9.892 on 304 degrees of freedom
## Multiple R-squared:  0.7616, Adjusted R-squared:  0.7585 
## F-statistic: 242.8 on 4 and 304 DF,  p-value: < 2.2e-16
```

For Model two, we added the Covid dummy variable, which makes this model a regression with Covid Dummy. After fitting the model here are the results.

Firstly, new_cpi and Covid coefficients are the only coefficients that are significant. The rest of the predictors, fuel_cpi and also fed_funds are not significant.

- new_cpi -> $p = 2 \times 10^{-16}$, way less than 0.05, significant.
- covid -> $p = 8.24 \times 10^{-5}$, way less than 0.05, significant.

After looking at the p-values and determining the significance, let’s focus on the Interpretations of the newly added dummy variable. After 2020, the used vehicle CPI level is about 10 points higher on average, even after accounting for other predictors.

From this model we can also see that the newly added dummy variable has some effects on the new_vehicle predictor, which shrinks it’s effect by a little; however, it still remains highly significant.

After analysis of the result we can see the goodness of fit in the summary table. From the summary table, we can see that adjusted $R^2 = 0.7585$. This means that the new model explains about 76 percent of the variance. Please note that SE also improved in this model to 9.892.

```
## [1] 2313.928
```

```
## [1] 2300.144
```

```
## [1] 2332.595
```

```
## [1] 2322.544
```

In summary, adding the Covid dummy variable has caused a clear structural shift. In terms of goodness of fit, model 2 fits better than model 1.

### Model Comparison

These are the evidence which led us to conclude that model 2 is a better model than model 1.

Model 1:

Adjusted R^2 = 0.7466, RE = 10.13, AIC = 2313.928, BIC = 2332.595

Model 2:

Adjusted R^2 = 0.7585, RE = 9.892, AIC = 2300.144, BIC = 2322.544

In comparison, Model 2 has higher adjusted R square, lower RE, lower AIC, BIC, which means that model 2 is the better model overall.

### Forecasting Results

```
##   t_future  new_cpi fuel_cpi fedfunds covid pred_used
## 1      310 165.4802 307.9374 1.555573     1  178.4958
## 2      311 165.5960 308.4276 1.552774     1  178.6176
## 3      312 165.7118 308.9179 1.549975     1  178.7395
## 4      313 165.8276 309.4082 1.547176     1  178.8614
## 5      314 165.9434 309.8984 1.544376     1  178.9833
```

After performing the methods that we mentioned we achieved the result table for the future new_cpi values. The model predicts a slow, steady increase in used car CPI over the next 5 months.

---

## Discussion

After disclosing results, it is time to have a discussion regarding them. In these two models that we fit, New Vehicle CPI has always been the strongest predictor, meanwhile Fuel prices & interest rates have small direct effects. We discovered in model 2 that COVID caused a structural upward shift, and after adding this variable, the final model explains rougly 76 % of variation.

This discovery led us to conclude that used cars became unusually expensive post-pandemic, which is quantified by our model (approximately 10 point level shift). The significant predictor New Vehicle CPI led us to conjecture that new car markets pressure spill over into used-car prices. This might also mean that there is a tight used-car supply chain as well as a high demand for used car. Meanwhile fuel prices and federal fund rates are irrelevant to the car market.

Please do note that there are some limitations to this project. Linear model may oversimplify the issue and ignore some other factors in play, and also predictors are are trending (non-stationary). Besides this, no lagged effects are being modeled, and no seasonality included. Lastly, forecasts rely on simple linear trends for predictors.

Some possible future improvements can include using SARIMA model to monitor differencing, seasonality, etc. It is possible that we can explore interactions or nonlinear models as well.

In conclusion, model 2 appears to be the best model that we have discovered. Forecast section has indicated stabilization in prediction values, which is what we needed to show.

# Appendix

```
library(tidyverse)
library(lubridate)
library(astsa)

used_raw <- read_csv("CUSR0000SETA02.csv") # Used Cars & Trucks CPI
```

```
## Rows: 873 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (1): CUSR0000SETA02
## date (1): observation_date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
new_raw  <- read_csv("CUSR0000SETA01.csv") # New Vehicles CPI
```

```
## Rows: 873 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (1): CUSR0000SETA01
## date (1): observation_date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
fuel_raw <- read_csv("CUSR0000SETB.csv")   # Motor Fuel CPI
```

```
## Rows: 705 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (1): CUSR0000SETB
## date (1): observation_date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
rate_raw <- read_csv("FEDFUNDS.csv")       # Federal Funds Effective Rate
```

```
## Rows: 856 Columns: 2
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl  (1): FEDFUNDS
## date (1): observation_date
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
# Quick previews
head(used_raw)
head(new_raw)
head(fuel_raw)
head(rate_raw)

used <- used_raw %>%
  mutate(observation_date = ymd(observation_date)) %>% # convert to date
  filter(observation_date >= ymd("2000-01-01")) %>% # filter for newer data
  rename(used_cpi = CUSR0000SETA02) # rename column

new <- new_raw %>%
  mutate(observation_date = ymd(observation_date)) %>%
  filter(observation_date >= ymd("2000-01-01")) %>%
  rename(new_cpi = CUSR0000SETA01)

fuel <- fuel_raw %>%
  mutate(observation_date = ymd(observation_date)) %>%
  filter(observation_date >= ymd("2000-01-01")) %>%
  rename(fuel_cpi = CUSR0000SETB)

rate <- rate_raw %>%
  mutate(observation_date = ymd(observation_date)) %>%
  filter(observation_date >= ymd("2000-01-01")) %>%
  rename(fedfunds = FEDFUNDS)

dat <- used %>%
  left_join(new,  by = "observation_date") %>%
  left_join(fuel, by = "observation_date") %>%
  left_join(rate, by = "observation_date") %>%
  arrange(observation_date)

dat <- dat %>%
  mutate(
    t = row_number(),
    covid = if_else(observation_date >= ymd("2020-01-01"), 1, 0)
  )

tsplot(dat$used_cpi, main = "Used Vehicles CPI (US)")
```

```
tsplot(dat$new_cpi, main = "New Vehicles CPI (US)")
```

```
tsplot(dat$fuel_cpi, main = "Motor Fuel CPI (US)")
```

```
tsplot(dat$fedfunds, main = "Federal Funds Rate (US)")
```

```
fit1 <- lm(used_cpi ~ new_cpi + fuel_cpi + fedfunds, data = dat)
summary(fit1)

fit2 <- lm(used_cpi ~ new_cpi + fuel_cpi + fedfunds + covid, data = dat)
summary(fit2)
```

```
AIC(fit1)
AIC(fit2)

BIC(fit1)
BIC(fit2)

trend_new  <- lm(new_cpi  ~ t, data = dat)
trend_fuel <- lm(fuel_cpi ~ t, data = dat)
trend_rate <- lm(fedfunds ~ t, data = dat)

future_t <- max(dat$t) + c(1, 2, 3, 4, 5)

future_new  <- predict(trend_new,  newdata = data.frame(t = future_t))
future_fuel <- predict(trend_fuel, newdata = data.frame(t = future_t))
future_rate <- predict(trend_rate, newdata = data.frame(t = future_t))

# Build future predictor dataframe
future_data <- data.frame(
  new_cpi  = future_new,
  fuel_cpi = future_fuel,
  fedfunds = future_rate,
  covid    = rep(1, length(future_new))  # still post-2020
)

# Predict future used vehicle CPI using best model (fit2)
pred_used <- predict(fit2, newdata = future_data)

# Combine for a neat table
future_results <- cbind(
  t_future = future_t,
  future_data,
  pred_used = pred_used
)

future_results
```

---