CAR MARKET ANALYSIS

Samuel Wu, Leo Yang, Royce Zhang

MOTIVATION

- Vehicles are important tools for transportation
- Vehicle prices are always varying in the U.S.
- Since Covid-19, vehicle prices surged and its effect remains strongly
- Understanding vehicle prices is helpful for customers
- Predicting future vehicle prices can help customers to be more informed

AGENDA - AN OVERVIEW OF OUR PRESENTATION

- Goal
- Data Source
- Methods
- Model 1 - 4
- Forecasting
- Conclusion
- Limitations
- Future Research

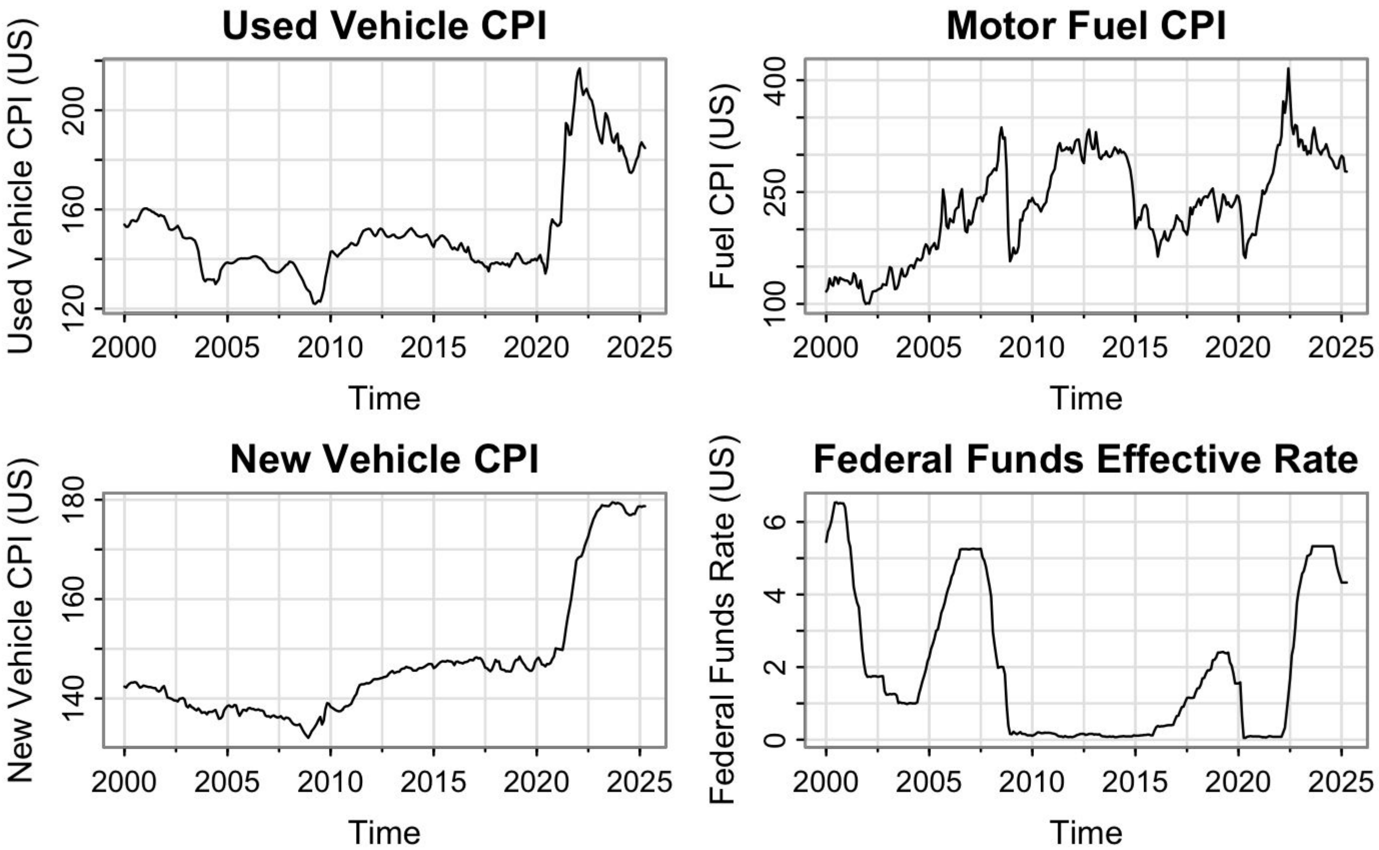
ANALYSIS A - REGRESSION ANALYSIS

1. Preliminary analysis
2. Regression model suggestions
3. Regression result analysis
4. Regression with autocorrelated errors
5. Final model selection
6. Forecast future 5 values

DATA SOURCE

- FRED : Federal Reserve Bank of St. Louis
  - Used Vehicle CPI
  - New Vehicle CPI
  - Fuel CPI
  - Federal Funds Effective Rates
- Joined on dates
- Covid-19: created as a dummy variable.
  - 0 means pre-covid time
  - 1 means post-covid time

Data Visualization



METHODS

1. Model 1: Linear Regression with COVID dummy
2. Model 2: Weighted Least Squares (WLS)
3. Model 3: Linear Regression with lagged predictors
4. Model 4: ARIMA (final model)

MODEL 1 - Linear Regression with Dummy Variable

```{r model 1, include=TRUE, echo=FALSE}
fit2 <- lm(used_cpi ~ new_cpi + fuel_cpi + fedfunds + covid, data = dat)
summary(fit2)
```

Model 1 Results

Summary Table

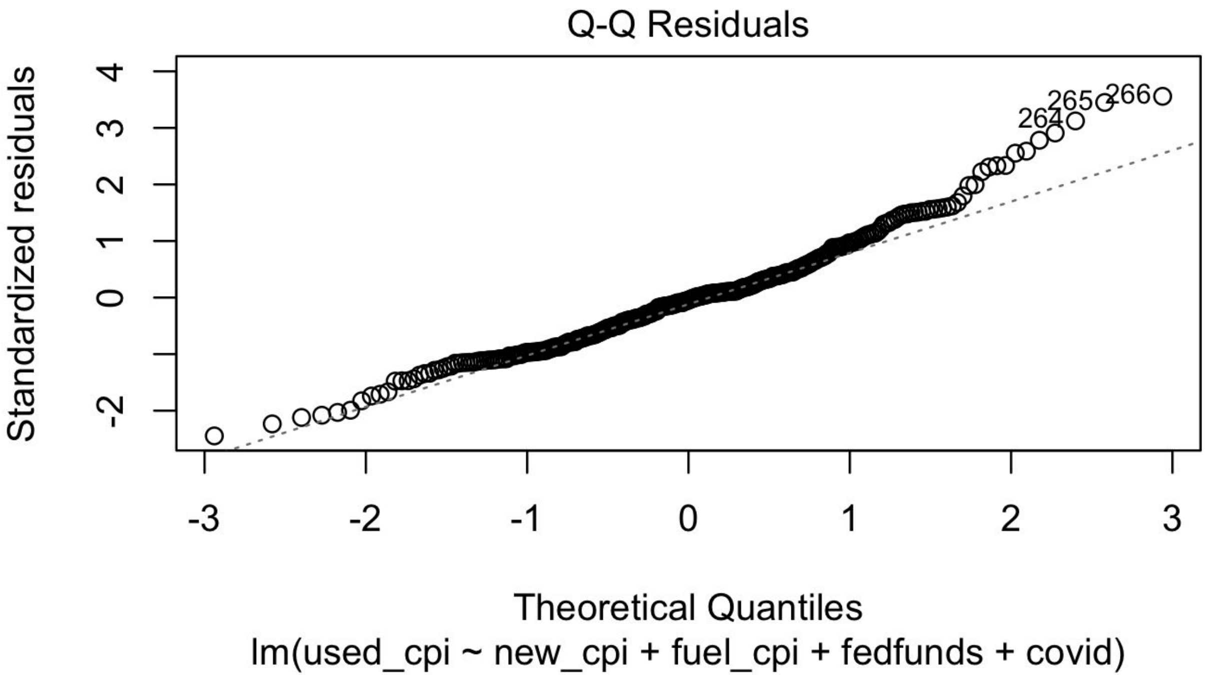
```
Call:
lm(formula = used_cpi ~ new_cpi + fuel_cpi + fedfunds + covid,
    data = dat)

Residuals:
    Min       1Q   Median       3Q      Max
-23.670   -7.127   -0.453    4.875   34.909

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.882950  11.990306  -0.657    0.511
new_cpi      1.056579   0.091388  11.561  < 2e-16 ***
fuel_cpi     0.005736   0.010505   0.546    0.585
fedfunds     0.212106   0.308909   0.687    0.493
covid        9.927836   2.515964   3.946 9.91e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.918 on 299 degrees of freedom
Multiple R-squared:  0.754,     Adjusted R-squared:  0.7507
F-statistic: 229.1 on 4 and 299 DF,  p-value: < 2.2e-16
```

Problem with Model 1 - Non-constant Variance in Residuals



Model 2: Linear Regression with Weighted Least Squares

```
```{r model 2, include=TRUE, echo=FALSE}
wt <- 1 / lm(abs(fit2$residuals) ~ fit2$fitted.values)$fitted.values^2

wls_model <- lm(used_cpi ~ new_cpi + fuel_cpi + fedfunds + covid,
                data = dat,
                weights = wt)

summary(wls_model)
```
```

Model 2 Results

Summary Table

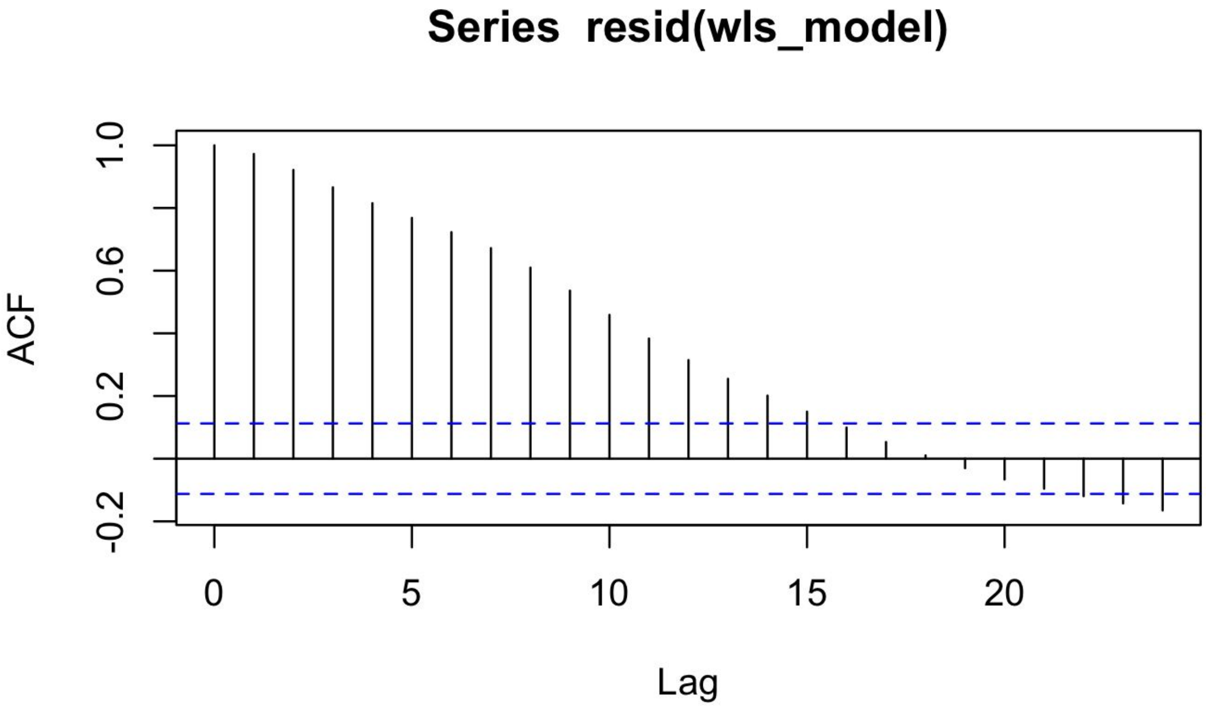
```
Call:
lm(formula = used_cpi ~ new_cpi + fuel_cpi + fedfunds + covid,
    data = dat, weights = wt)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-2.3925 -0.9062 -0.0985  0.8912  3.3071

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -28.026577  11.735751  -2.388  0.01755 *
new_cpi       1.215372   0.086073  14.120  < 2e-16 ***
fuel_cpi     -0.009485   0.008125  -1.167  0.24402
fedfunds      0.815244   0.252607   3.227  0.00139 **
covid         6.034583   2.511593   2.403  0.01688 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.214 on 299 degrees of freedom
Multiple R-squared:  0.6489,	Adjusted R-squared:  0.6442
F-statistic: 138.1 on 4 and 299 DF,  p-value: < 2.2e-16
```

Problem with Model 2 - ACF Shows Past Correlation



MODEL 3 - Linear Regression with Lagged Term

```
```{r, echo=FALSE}
dat <- dat %>%
mutate(
new_cpi_lag1 = lag(new_cpi, 1),
fuel_cpi_lag3 = lag(fuel_cpi, 3),
fedfunds_lag12 = lag(fedfunds, 12)
)

fit3 <- lm(used_cpi ~ new_cpi + new_cpi_lag1 + fuel_cpi + fuel_cpi_lag3+ fedfunds + fedfunds_lag12, data = dat)

summary(fit3)
```
```

Model 3 Results

Summary Table

```
Call:
lm(formula = used_cpi ~ new_cpi + new_cpi_lag1 + fuel_cpi + fuel_cpi_lag3 +
    fedfunds + fedfunds_lag12, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-20.290  -6.381  -0.543   5.159  36.020

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    -45.65844    7.36980  -6.195 2.03e-09 ***
new_cpi          7.05699    0.98794   7.143 7.63e-12 ***
new_cpi_lag1    -5.73826    0.99988  -5.739 2.43e-08 ***
fuel_cpi         0.04840    0.02183   2.217   0.0274 *
fuel_cpi_lag3   -0.04212    0.02139  -1.969   0.0499 *
fedfunds        -0.77780    0.45426  -1.712   0.0879 .
fedfunds_lag12   0.96717    0.41370   2.338   0.0201 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.396 on 285 degrees of freedom
  (12 observations deleted due to missingness)
Multiple R-squared:  0.7889,    Adjusted R-squared:  0.7845
F-statistic: 177.5 on 6 and 285 DF,  p-value: < 2.2e-16
```

MODEL 4 - ARIMA Model

```{r}
fit_arima_best <- Arima(dat$used_cpi, order = c(0,1,2))
print(summary(fit_arima_best))
```

Model 4 Results

Summary Table

```
=== Fit possible ARIMA models ===

ARIMA Model Comparison:

*** Most Optimal ARIMA Model: ARIMA(0,1,2) ***
Series: dat$used_cpi
ARIMA(0,1,2)

Coefficients:
         ma1     ma2
      0.6444  0.3590
s.e.  0.0537  0.0541

sigma^2 = 3.445:  log likelihood = -616.57
AIC=1239.14   AICc=1239.22   BIC=1250.29

Training set error measures:
                     ME     RMSE      MAE       MPE     MAPE      MASE         ACF1
Training set 0.05120644 1.846776 1.065576 0.0303407 0.671188 0.7972187 -0.002600269
```

MODEL COMPARISON

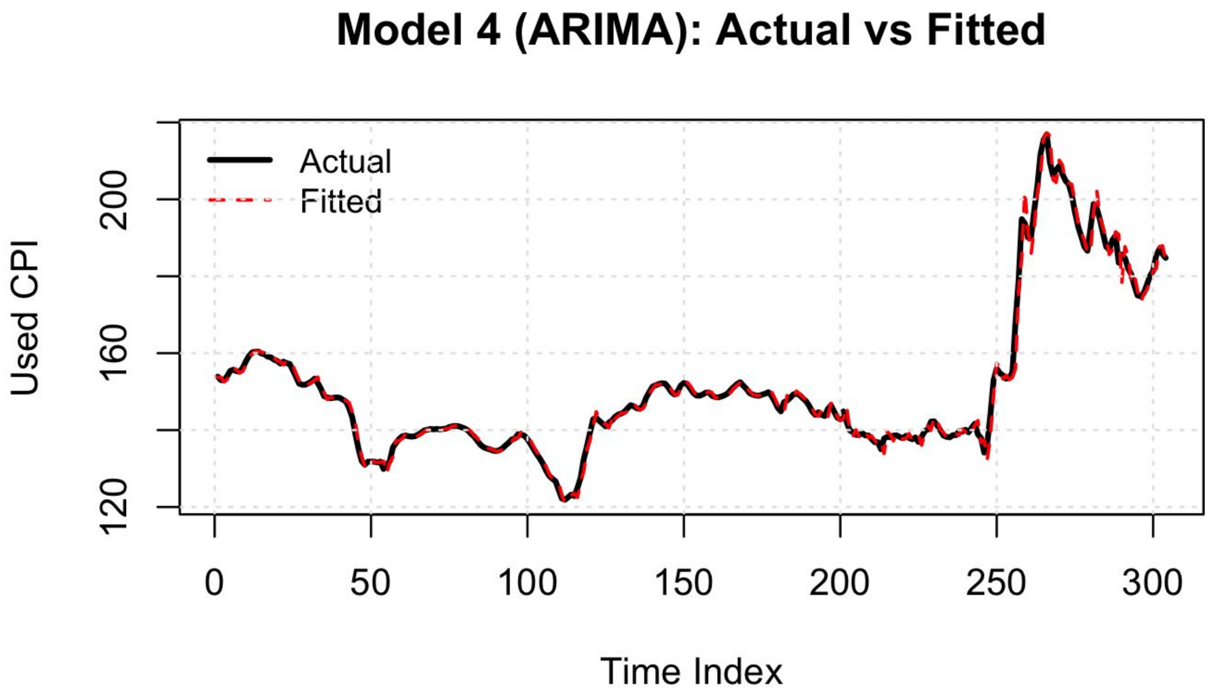
Model	Adjusted R^2	AIC	BIC	RMSE
1. Regression with Dummy	0.7507	2264.617	2286.920	9.836
2. WLS	0.644	2190.484	2212.786	10.034
3. Lagged Term	0.793	2193.214	2230.285	9.081
4. ARIMA(0,1,2)	-	1239.14	1250.29	3.445

FORECASTING

	Point Forecast <dbl>	Lo 80 <dbl>	Hi 80 <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
305	184.4098	182.0313	186.7883	180.7722	188.0474
306	184.6285	180.0509	189.2062	177.6277	191.6294
307	184.6285	178.0208	191.2362	174.5229	194.7341
308	184.6285	176.4818	192.7752	172.1692	197.0878
309	184.6285	175.1906	194.0665	170.1944	199.0627

5 rows

Forecasting - 5 month ahead forecast using ARIMA(0,1,2)



CONCLUSION

- Model 4 is the best model overall
- It has the lowest AIC, BIC, and RMSE for all models
- It successfully eliminates the issue of auto-correlation and partial autocorrelation
- The prediction results closely aligns with the real values

LIMITATIONS

- Variables are trending & non-stationary
- No guarantees of causal interpretation
- Structures other than Covid were ignored
- Forecasting approach is simplified

FUTURE RESEARCH

- Extend the model into ARIMAX or SARIMAX
- Explore lag structures more formally
- Include more economic variables
- Use machine learning for nonlinearity

THANK YOU

CITATIONS

- **Consumer Price Index for All Urban Consumers: Used Cars and Trucks in U.S. City Average** (CUSR0000SETA02)
  - https://fred.stlouisfed.org/series/CUSR0000SETA02
- **Consumer Price Index for All Urban Consumers: Fuel Oil and Other Fuels in U.S. City Average** (CUSR0000SEHE)
  - https://fred.stlouisfed.org/series/CUSR0000SEHE
- **Consumer Price Index for All Urban Consumers: New Vehicles in U.S. City Average** (CUUR0000SETA01)
  - https://fred.stlouisfed.org/series/CUUR0000SETA01
- **Federal Funds Effective Rate** (FEDFUNDS)
  - https://fred.stlouisfed.org/series/FEDFUNDS