

# Traffic, Population & Employment as Predictors of US GDP

By Kamron Afshar & Shishir Kumar

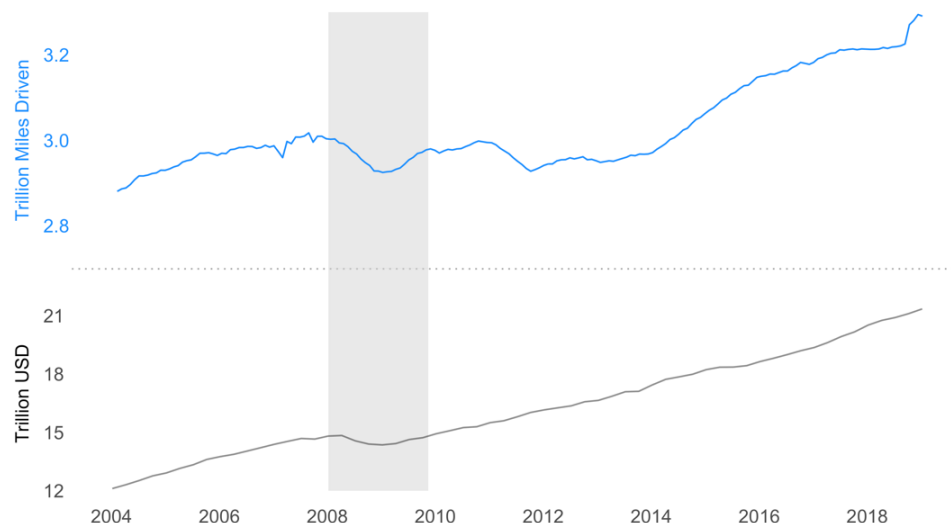
1. **Dataset:** 60 values corresponding with 3-month calendar quarters
  - a. Response Variable
    - i. GDP: US Gross Domestic Product in Billion US Dollars
  - b. Predictors
    - i. 12 month rolling miles driven across us, averaged at the quarter level
    - ii. % total mileage contribution in the North East US
    - iii. % total mileage contribution in the North Central US
    - iv. % total mileage contribution in the South Atlantic US
    - v. % total mileage contribution in the West US
    - vi. % total mileage contribution from Urban Arterial Roads
    - vii. US Population
    - viii. % US Employment
2. **Research Problem:** Living in the Bay Area during the “Great Recession of 2008” Kamron noticed a decrease in the traffic levels, which later increased greatly as the economy recovered. Furthermore, there is some logical credence to the notion that traffic, population, and employment can generate or indicate a strong economy.

We determined based on the graphs below there is some relationship between traffic and the US GDP which led us to further analyze this relationship.

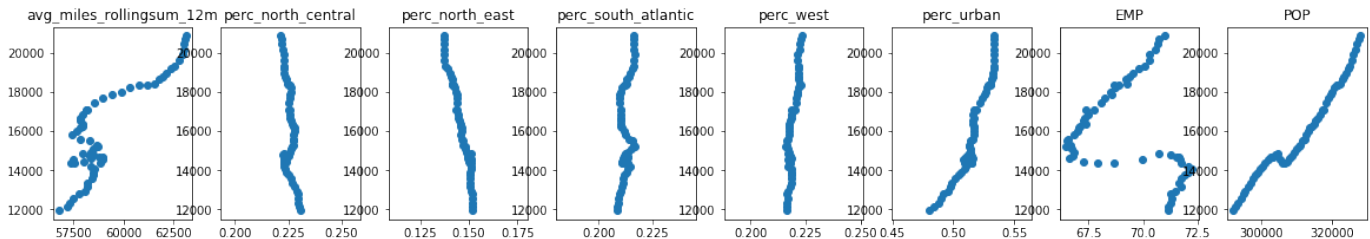
We chose to attempt a Multiple Linear Regression model to determine which combination of the variables can best predict the US GDP. Given the nature of the data is clear there is some autocorrelation/ time related effect, but we will attempt to use MLR; a Time Series Analysis would probably be the most appropriate.

3. **Exploratory Analysis:** The graph below is what encouraged us to examine this trend in the first place. It shows 12 month rolling sum of the overall vehicular miles driven in the US every year, contrasted with yearly GDP. We can see that the movement in GDP is almost mirrored by miles data. It is important to note that there is a dip during the recession of 2008 which will have an affect across multiple variables in our model.

**Traffic Volume Moves with US GDP**

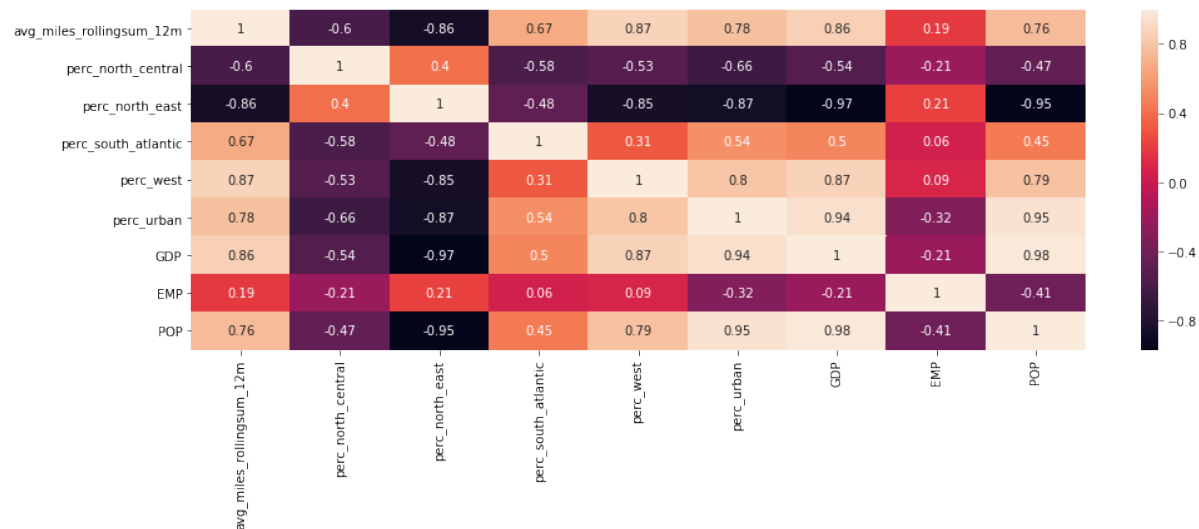


Based on the following scatter plots, there does appear to be strong relationships between some of these variables although there is an irregularity within the plots for Employment Rate, Rolling Sum of miles, and Population. We suspect this is due the effect of the recession on these variables. It also might be worthwhile to consider rescaling the data in some way based on the shear slope of the relationships indicated in the scatter plots.



Based on these scatter plots we felt this analysis can be fruitful. That being said, we must consider that the direction of relationship we are assuming here might be reversed. Another analysis might treat our predictor variables as responses to GDP instead of GDP as the response.

Also, our variables for percentage contribution of miles driven are interrelated which will probably lead to correlation between these variables. The correlation heat map below confirms many of our predictors have intercorrelation.



**4&5: Regression Analysis & Model Selection:** We started with an MLS model with all of our variables included which produced the regression summary below:

OLS Regression Results

Dep. Variable:

GDP

R-squared:

0.997

Model:

OLS

Adj. R-squared:

0.997

Method:

Least Squares

F-statistic:

2505.

Date:

Sun, 13 Oct 2019

Prob (F-statistic):

2.26e-63

Time:

03:30:51

Log-Likelihood:

-372.29

No. Observations:

60

AIC:

762.6

Df Residuals:

51

BIC:

781.4

Df Model:

8

Covariance Type:

nonrobust

coef

std err

t

P>|t|

[0.025

0.975]

Intercept

-1.011e+05

1.66e+04

-6.075

0.000

-1.34e+05

-6.77e+04

avg\_miles\_rollingsum\_12m

-0.1009

0.048

-2.120

0.039

-0.196

-0.005

perc\_north\_central

1.061e+04

1.85e+04

0.572

0.570

-2.66e+04

4.78e+04

perc\_north\_east

5112.3242

3.05e+04

0.168

0.867

-5.6e+04

6.63e+04

perc\_south\_atlantic

5.525e+04

1.96e+04

2.822

0.007

1.59e+04

9.46e+04

perc\_west

6.735e+04

3.37e+04

1.999

0.051

-303.523

1.35e+05

perc\_urban

-1.833e+04

9329.298

-1.964

0.055

-3.71e+04

403.630

EMP

297.2237

27.069

10.980

0.000

242.880

351.568

POP

0.2645

0.021

12.881

0.000

0.223

0.306

Omnibus:

0.424

Durbin-Watson:

0.811

Prob(Omnibus):

0.809

Jarque-Bera (JB):

0.575

Skew:

-0.159

Prob(JB):

0.750

Kurtosis:

2.642

Cond. No.

8.01e+08

This regression summary output helped us determine at a surface-level which variable were not helpful in prediction. R Squared value is very high but that is probably due to the correlation between the variables. Just based on the P values we chose to remove % North Central and % North East. Although the p value for % Urban and % West variables are over 5%, they are very close so we kept them in the model.

After removing these variables, we noted a decrease in AIC & BIC while maintaining the same R Squared. The p values for % contribution West and Urban decreased to below the threshold which led to us keeping these variables.

An analysis of VIF indicated there is strong multicollinearity in our data unfortunately. This is to be expected based on our understanding of these variables. The relationship between any two of these variables would provide ample opportunity for analysis on their own. Given this however we chose to proceed with the given variables in our second model: Miles, % South Atlantic, % West, % Urban, Employment Rate and Population.

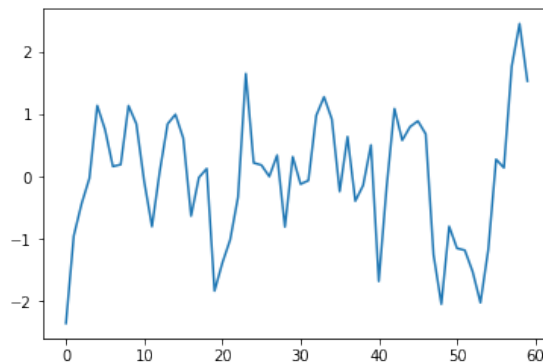
OLS Regression Results							
Dep. Variable:	GDP		R-squared:	0.997			
Model:	OLS		Adj. R-squared:	0.997			
Method:	Least Squares		F-statistic:	3448.			
Date:	Sun, 13 Oct 2019		Prob (F-statistic):	7.72e-67			
Time:	03:30:54		Log-Likelihood:	-372.49			
No. Observations:	60		AIC:	759.0			
Df Residuals:	53		BIC:	773.6			
Df Model:	6						
Covariance Type:	nonrobust						
		coef	std err	t	P> t	[0.025	0.975]
Intercept		-9.522e+04	7259.016	-13.118	0.000	-1.1e+05	-8.07e+04
avg_miles_rollingsum_12m		-0.0959	0.045	-2.136	0.037	-0.186	-0.006
perc_south_atlantic		5.21e+04	1.79e+04	2.914	0.005	1.62e+04	8.8e+04
perc_west		6.689e+04	3.24e+04	2.068	0.044	1999.845	1.32e+05
perc_urban		-2.034e+04	5129.333	-3.964	0.000	-3.06e+04	-1e+04
EMP		288.0799	19.587	14.708	0.000	248.794	327.366
POP		0.2626	0.008	33.602	0.000	0.247	0.278
Omnibus:	1.073	Durbin-Watson:	0.803				
Prob(Omnibus):	0.585	Jarque-Bera (JB):	1.100				
Skew:	-0.216	Prob(JB):	0.577				
Kurtosis:	2.497	Cond. No.	6.99e+08				

## 6. Model Diagnosis:

a. First we analyzed the heteroskedasticity. LM and F probability values were very close to 5% but slightly over. There is likely some heteroskedasticity but given these results arguably the assumption of constant variance is safe.

b. Next we tested Auto correlation. P values for both LM and F indicated there is definitely autocorrelation. This was our assumption from the beginning, but we chose to proceed to learn more about the data. This confirms that we should be using a time series type analysis here in place of MLR.

c. Based on the plot of externally studentized residuals below, the residuals appear to be randomized around 0 so that assumption is satisfied.



d. based on the DFFITS the first value, the 24<sup>th</sup> value and the second to last value appear to be influential. We can't seem to explain why the first and last would be influential so those weren't removed. The 24<sup>th</sup> however is related to the recession, which is somewhat expected. We believe we should keep this data point because it is an important element in analyzing the relationships between these predictors and GDP. Perhaps another variable to indicated where the economy is in recession or not would be valuable to account for a change in the relationships during economic downturns.

**7. Model of Choice:** We performed a general linear F test between the reduced and full mode. The p-value comes out to be much greater than 0.05, hence we cannot reject the null hypothesis and accept the reduced model with parameters as mileage, % South Atlantic, % West, % Urban, Employment Rate and Population. The coefficients are a bit hard to understand given some are negative despite a seeming positive correlation in the individual scatter plots. Perhaps the negative coefficient is accounting for the multicollinearity in a way. We expected each variable to have a positive coefficient.

**8. Summary:** While we were able to generate a model with a high R Squared value, there is obviously a lot of underlying interactions that render this model less than ideal for accurate predictions. Most importantly, this would be a more apt example of a time series analysis problem instead of MLR. That being said we were able to prove there is

some relationship between these variables as expected, and in the case of some of the variables a visibly strong relationship. The expertise of an economist would be helpful in determining other potential variables we should include as well as potential transformation to these variables/response to account for the interrelationship and economic cycles. Some of our variables are “adjusted values” which rely on some analysis that was not clear to us to smooth out or improve accuracy. To take this model further we would need to understand how the adjustments work and apply them across all our variables to improve accuracy.