

Predicting Individual Fuel Economy

2011-01-0618

Published
04/12/2011

Zhenhong Lin and David Greene
Oak Ridge National Laboratory

doi:10.4271/2011-01-0618

ABSTRACT

To make informed decisions about travel and vehicle purchase, consumers need unbiased and accurate information of the fuel economy they will actually obtain. In the past, the EPA fuel economy estimates based on its 1984 rules have been widely criticized for overestimating on-road fuel economy. In 2008, EPA adopted a new estimation rule. This study compares the usefulness of the EPA's 1984 and 2008 estimates based on their prediction bias and accuracy and attempts to improve the prediction of on-road fuel economies based on consumer and vehicle attributes. We examine the usefulness of the EPA fuel economy estimates using a large sample of self-reported on-road fuel economy data and develop an Individualized Model for more accurately predicting an individual driver's on-road fuel economy based on easily determined vehicle and driver attributes. Accuracy rather than bias appears to have limited the usefulness of the EPA 1984 estimates in predicting on-road MPG. The EPA 2008 estimates appear to be equally inaccurate and substantially more biased relative to the self-reported data. Furthermore, the 2008 estimates exhibit an underestimation bias that increases with increasing fuel economy, suggesting that the new numbers will tend to underestimate the real-world benefits of fuel economy and emissions standards. By including several simple driver and vehicle attributes, the Individualized Model reduces the unexplained variance by over 55% and the standard error by 33% based on an independent test sample. The additional explanatory variables can be easily provided by the individuals.

INTRODUCTION

The oil price shock of 1974 led to the requirement to label new light-duty vehicles with lab-based estimates of city and highway fuel economy. In response to consumer complaints about the lab-based estimates, the U.S. Environmental

Protection Agency (EPA) carried out a statistical analysis of a non-random sample of on-road fuel economy estimates from tens of thousands of vehicles, and in 1984 adopted downward adjustment factors of 10% and 22% respectively for the city and highway lab-based estimates [1-2]. The EPA combined fuel economy was calculated by harmonically averaging the city and highway estimates weighted 55% and 45%, respectively. Although the number of consumer complaints declined significantly after EPA adopted the 1984 estimates [1], the perception that the 1984 estimates still overestimated on-road fuel economy remained [3, 4, 5]. In 2008, EPA adopted a new estimation method with the objective of "... giving consumers a more accurate estimate of the fuel economy they can achieve in the real-world" [1]. No large-sample empirical study is available that compares the usefulness of the 1984 and 2008 estimates to consumers.

In this study, we use on-road fuel economy data voluntarily submitted by users to www.fueleconomy.gov to examine the usefulness of the EPA's 1984 and 2008 fuel economy estimates based on the ability to consistently and accurately predict the reported on-road fuel economy numbers. We then use the same data to specify a statistical model that more accurately predicts reported on-road fuel economy based on a few consumer-specific and vehicle-specific attributes. More accurate individualized fuel economy estimates should increase the value of fuel economy information to consumers, possibly leading to more efficient choices of vehicle technologies [6] and better informed travel behavior. Like the data used to generate the adjustment factors for the 1984 estimates, our data sample is a convenience sample, not a scientific random sample. Inferences based on convenience samples cannot be considered scientifically definitive.

This paper is organized as follows. The next section describes the three data samples and three models used to support the analysis. In the Results and Discussions section, we first

compare the usefulness of the 1984 and 2008 estimates and then further examine the bias of the two EPA estimates. We then show the limitation of adjustment by vehicle technology without additional driver and vehicle attributes. After specifying the Individualized Model, we compare its on-road fuel economy predictions to the EPA estimates. Finally, we use the Individualized Model to assess the potential impact of driver and vehicle attributes on on-road fuel economy.

DATA AND METHODS

DATA SOURCE AND STRUCTURE

Data were obtained from the “Your MPG” system of www.fueleconomy.gov, a website of the U.S. Department of Energy (DOE) and EPA. The system allows individuals to voluntarily submit their own on-road fuel economy estimates (referred to as “YourMPG estimates” in text and “YourMPG” in equations) based on dashboard read-outs, self calculation, trip distances with fuel purchase records, odometer readings with fuel purchase records, or other methods. Individuals first identify the make, model, engine, transmission and model year of their vehicle by selecting it from www.fueleconomy.gov's online database. When a vehicle model is selected, several other vehicle attributes are automatically retrieved from the system, including the city, highway and combined fuel economy ratings by each of the 1984 and 2008 EPA rules. Throughout the paper, we use the terms “1984 estimates” (“EPAMPG₁₉₈₄” in equations) and “2008 estimates” (“EPAMPG₂₀₀₈” in equations) to represent the EPA's published combined fuel economy ratings resulting from the lab testing, city/highway share and adjustment prescribed by the EPA's 1984 and 2008 rules, respectively [1]. Individuals may also report the following information associated with the YourMPG estimate: 1) the driving style on a scale of 0 to 5 (5 = most aggressive), 2) the self-estimated distance share of stop-and-go driving (as opposed to highway driving), 3) the state of residence, and 4) the time period. The time period and the state were used to retrieve the associated historical ambient temperatures. Each data record represents one vehicle. If the method is reporting multiple trip distances or odometer readings with fuel purchase records, the data record is based on the weighted harmonic mean of the submissions.

MODELS AND SAMPLES

Three statistical models and three samples were used for different tests and their relationship is illustrated in Figure 1. First, the bias and accuracy of the EPA 1984 and 2008 MPG estimates were measured against the YourMPG estimates using the Naive Model (1) with Sample A. The Naive Model (1), where i ($i \in \text{Sample A}$) represents a vehicle record in Sample A and k ($k=1984$ or 2008) indicates which EPA ratings is of concern, assumes that consumers take the EPA estimates at face value. Sample A comprises 31227 records

that were submitted between September 2004 and November 2009 and is used to estimate the distribution of prediction errors (ϵ_{ki}) by the EPA estimates.

$$\text{Naive Model: YourMPG}_i = 0 + 1 \cdot \text{EPAMPG}_{ki} + \epsilon_{ki} \quad (1)$$

$$\text{Linear Model: YourMPG}_i = a_{kt} + b_{kt} \cdot \text{EPAMPG}_{ki} + \delta_{kti} \quad (2)$$

$$\text{Individualized Model: YourMPG}_j = c^T A_j + \theta_j \quad (3)$$

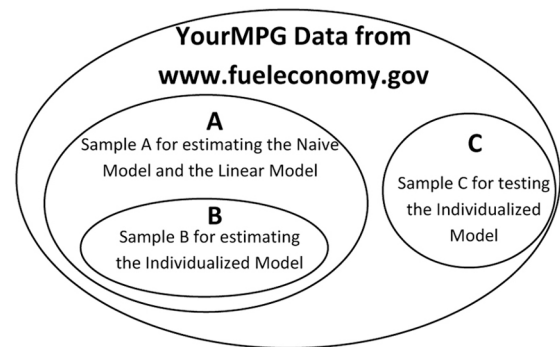


Figure 1. Relationship of Samples and Models

The bias of the EPA estimates was further analyzed using Sample A and the Linear Model (2), where t indicates each of the 5 vehicle technologies (gasoline, diesel, hybrid, ethanol FFV, and compressed natural gas (CNG)) and δ_{kti} is the residual assumed to following independent normal distribution $N(0, \sigma_{kt}^2)$. The Linear Model (2), estimated via ordinary least squares (OLS) regression, measures the constant and relative biases of the two EPA estimates EPAMPG_k, as well as measuring technology-specific biases.

The Individualized Model (3) (where A_j is the vector of driver and vehicle attributes associated with vehicle j ($j \in \text{Sample B}$), c^T is the vector of regression coefficients, and θ_j is the residual assumed to follow an independent normal distribution $N(0, \sigma^2)$) is intended to create a more accurate predictor of on-road fuel economy. It was estimated using Sample B, a subset of Sample A. In generating Sample B, each record was required to have valid values for all the variables in the Individualized Model and its YourMPG estimate was required to fall within 3 standard errors of EPAMPG₁₉₈₄. The standard errors are technology specific and were calculated based on the error term (YourMPG-EPAMPG₁₉₈₄). The estimated standard errors for the 5 vehicle technologies are shown in Table 1. The 3-standard-error test identified 324 records (~1%) in Sample A as outliers. Among them, only 79 records passed the valid value test, so the 3-standard-error test removed only 79 outliers. In

the remaining 8849 records, there are only 10 records where the technology type of the vehicle is CNG. Due to such a small representation, the 10 CNG records were also removed. This resulted in 8839 records in Sample B.

Table 1. Technology Specific Standard Deviation

Descriptive Statistics of (YourMPG-EPAMPG ₁₉₈₄)					
	Gasoline	Diesel	Hybrid	CNG	FFV
Mean	0.640	3.164	-4.357	4.319	0.015
Median	0.500	3.000	-4.313	3.915	0.000
Maximum	31.610	23.000	35.600	16.000	12.822
Minimum	-22.675	-11.505	-43.000	-1.000	-8.471
Std. Dev.	4.029	4.917	6.131	3.345	3.221
Skewness	0.283	0.490	0.369	0.843	0.503
Kurtosis	4.816	4.071	7.571	4.392	4.094
Observations	27895	920	1900	51	461

Finally, the Individualized Model estimated based on Sample B was tested on the independent Sample C, consisting of 1082 records submitted between November 2009 and May 2010, with no missing values for the explanatory variables and within 3 standard errors of EPAMPG₁₉₈₄.

The analysis was conducted in units of miles-per-gallon (MPG) rather than gallons-per-mile (GPM), although it is recognized that the MPG metric can lead consumers to incorrectly estimate fuel savings when comparing vehicles [7]. As Figure 2 shows, the relationship between the 1984 estimates and the YourMPG estimates appears linear in MPG space (for a given technology) but less so in GPM space (not plotted).

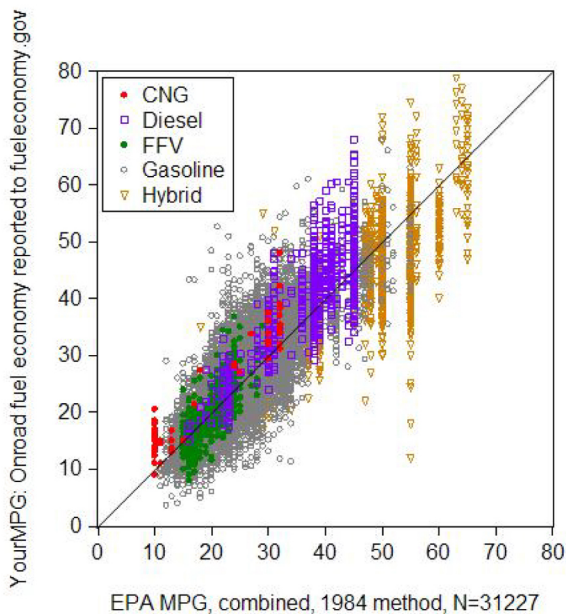


Figure 2. EPA MPG 1984 vs. YourMPG

On scatter plots of Sample A, both the EPA 1984 and 2008 estimates exhibit an approximately linear correlation with

YourMPG, the self-reported on-road fuel economy (Figure 2 and Figure 3). While the 1984 estimates appear to scatter around the line $y=x$, the 2008 estimates appear to result from a horizontal contraction of the 1984 estimates around the low efficiency points. This suggests that the 2008 estimates have a larger downward adjustment from the 1984 estimates for more efficient vehicles. Both estimates exhibit vehicle technology-specific variances. This suggests that any statistical adjustment of fuel economy estimate may need to be differentiated by vehicle technology.

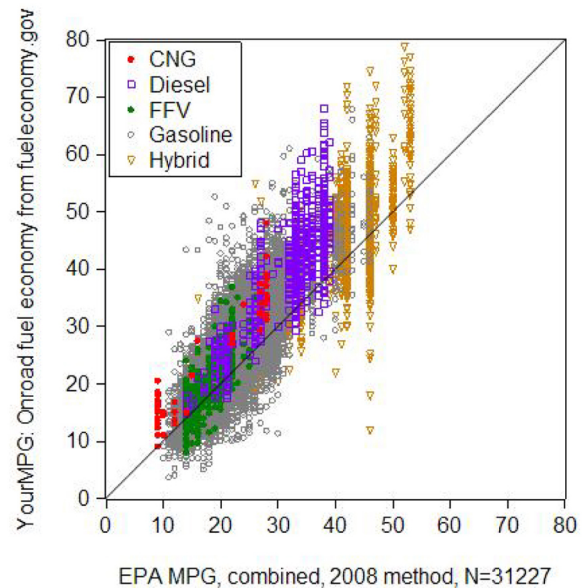


Figure 3. EPA MPG 2008 vs. YourMPG

SAMPLE BIAS AND DATA QUALITY

Any data set of this type will contain measurement errors, despite a variety of automated checks performed on each entry. Fuel purchase and odometer records entered by participants can be rigorously checked for errors but other types of estimates can only be screened for outliers. Different measurement methods may result in varying biases. This is addressed by interacting the measurement method with both the constant and slope terms in the Individualized Model. Respondents report their driving style (on a scale of 0-5) and stop-and-go (as opposed to free-flowing) driving shares (0-100%), but no instructions are given on how to measure them. Therefore, the rating subjectivity will also contribute to the estimation variance.

Very few of the 2008 mpg estimates in the dataset were actually obtained using the new 5-cycle test methods, which include 3 additional test cycles and are intended to incorporate additional factors like higher speeds, more aggressive driving, the use of air conditioner and cold temperature, as described in the final regulation [8]. Relatively few vehicles have been tested on all five cycles.

Instead, the vast majority of the 2008 mpg estimates are derived from the city and highway test estimates using equations developed by EPA to approximate the 5-cycle tests. Auto manufacturers are not required to use the vehicle-specific 5-cycle approach until the model year 2011. For model years 2008-2010, a manufacturer can choose to use either approach. At the time of writing, we have only 11 data records with the 2008 estimates from the vehicle-specific 5-cycle approach. We used the traditional city/highway weighting factors of 55/45 throughout this analysis, although EPA in its fleet-wide fuel economy estimation report has begun to adopt a 43/57 city/highway weighting [9].

Overall, the three samples have similar data distributions by estimation method, vehicle technology, vehicle class, transmission, stop-and-go share and driving style (Table 2). Take Sample A as an example. It contains a reasonably representative distribution of vehicles categories: 65% are passenger cars and 35% are light trucks. Among passenger cars, compacts are the most common (20%), followed by midsize (17%) and subcompact (13%). Among light trucks, SUVs (17%) and pickups (10%) are the most common. Conventional gasoline vehicles (89%) dominate the sample. Hybrids (6%) and diesels (3%) are overrepresented as compared to the vehicle population on the road. User self-calculation (46%) is the most common source of data. Trip meter and odometer methods, which use the Your MPG website software for calculation, make up 38% of the data. Nearly 14% of Sample A comes from onboard dash readouts.

Because the dataset is not based on random sampling, it might contain a self-selection bias towards drivers more interested in fuel economy and therefore more likely to achieve higher fuel economy than the population of U.S. drivers. However, the distribution of self-reported driving style in Table 2 is slightly skewed towards aggressive driving styles, suggesting that the respondents do not consider themselves to be above average in this respect. As shown by the Linear Model specifications by vehicle technology in Table 3, a linear regression of the YourMPG estimates against the 1984 estimates on conventional gasoline vehicles yielded an intercept of 0.07 ($p=0.4826$) and a slope coefficient of 1.02 ($p<0.001$), indicating a 2% underestimation bias. This regression result is consistent with a previous paper [10] that used a smaller sample of the YourMPG data 4 years ago. There are two possible explanations for the consistency of the 1984 and YourMPG estimates: 1) the YourMPG dataset is not substantially biased and the 1984 estimates are still approximately unbiased, or 2) almost exactly the same amount of bias is present in the original 1984 data set and the YourMPG dataset (both 4 years ago and today). While both are possible, explanation 2) requires a remarkable coincidence. Therefore, we lean toward the first explanation. However, because the 1984 sample is also a convenience sample and there have been changes in vehicle technology and traffic behavior since then, we cannot

rule out the possibility of driver bias. We leave this to the reader's judgement.

Table 2. Distribution of Sample A, B, C

Sample	A (N=31227)		B (N=8839, B⊂A)		C (N=1082, C∩A=∅)	
	Freq.	100%	Freq.	100%	Freq.	100%
Method¹						
Calculated	14444	46.3	3174	35.9	474	43.8
Trip Meter	6285	20.1	2434	27.5	249	23.0
Odometer	5574	17.9	2024	22.9	228	21.1
Dash	4323	13.8	1054	11.9	124	11.5
Mixed	303	1.0	96	1.1	0	0.0
Guess	106	0.3	28	0.3	3	0.3
Missing	147	0.5	16	0.2	4	0.4
Other	45	0.1	13	0.2	0	0.0
Total	31227	100.0	8839	100.0	1082	100.0
Technology						
Gasoline	27895	89.3	7981	90.3	936	86.5
Hybrid	1900	6.1	512	5.8	86	8.0
Diesel	920	3.0	204	2.3	40	3.7
Ethanol FFV	461	1.5	142	1.6	20	1.9
CNG	51	0.2	0	0.0	0	0.0
Total	31227	100.0	8839	100.0	1082	100.0
Vehicle Class						
Subcompact	3944	12.6	1167	13.2	123	11.4
Compact	6319	20.2	1815	20.5	222	20.5
Midsize	5329	17.1	1549	17.5	219	20.2
Large	1624	5.2	455	5.2	56	5.2
Wagons	2056	6.6	603	6.8	74	6.8
Two Seaters	854	2.7	261	3.0	50	4.6
SUV	5228	16.7	1443	16.3	180	16.6
Pickup	3239	10.4	878	9.9	110	10.2
Minivan	1171	3.8	320	3.6	35	3.2
Vans	248	0.8	62	0.7	0	0.0
Spl. Purpose	1215	3.9	286	3.2	13	1.2
Total	31227	100.0	8839	100.0	1082	100.0
Transmission Type						
Automatic	22907	73.4	6483	73.4	801	74.0
Manual	8320	26.6	2356	26.7	281	26.0
Total	31227	100.0	8839	100.0	1082	100.0
Distance Share of Stop and Go Driving						
[0, 0.2)	4233	13.6	1232	13.9	145	13.4
[0.2, 0.4)	8741	28.0	2723	30.8	309	28.6
[0.4, 0.6)	9422	30.2	2882	32.6	365	33.7
[0.6, 0.8)	2981	9.6	1070	12.1	131	12.1
[0.8, 1]	3356	10.7	932	10.5	132	12.2
missing	2494	8.0	0	0.0	0	0.0
Total	31227	100.0	8839	100.0	1082	100.0
Driving Style³						
[0,1)	226	0.7	193	2.2	20	1.8
[1,2)	1982	6.3	1715	19.4	182	16.8
[2,3)	2810	9.0	2408	27.2	313	28.9
[3,4)	3445	11.0	2968	33.6	369	34.1
[4,5]	1788	5.7	1555	17.6	198	18.3
missing or invalid	20976	67.2	0	0.0	0	0.0
Total	31227	100.0	8839	100.0	1082	100.0

1. "Calculated": YourMPG is mathematically calculated by users. "Odometer": users provide odometer readings and fuel purchases and let the website software calculate YourMPG. "Trip Meter": users provide miles travelled between fill-ups and fuel purchases and let the website software calculate YourMPG. "Dash": YourMPG comes from onboard automated fuel economy calculators. "Mixed": multiple methods are used to estimate YourMPG. "Guess": users state that YourMPG is based on guessing. "Other": none of the above, e.g. the user cites an estimate from a source or friend.

2. "CNG" includes bi-fuels vehicles where one fuel is CNG. "Gasoline" does not include gasoline-powered hybrid vehicles, which are included in "Hybrid".

3. The highest level 5 refers to the most aggressive driving style.