AbhijitAmin
JamisonAndaluz
RobinAzzam
NirmalRajan
RichardSoni

Cars Going Green

Abstract:

Carbon Footprint is a hot topic in our current society. Cars play a key role in Greenhouse Gas emissions. If the amount of toxins being launched into our atmosphere does not diminish our children will suffer. We attempted to find the elements in a motor vehicle that affects its GHGS (defined below). We found that there are variables that affect this, and that some can be used to predict the GHGS of a vehicle. The model that best explains the significant effect on the Greenhouse Gas Score of a vehicle, and the one we used for our predictions, uses its combined mileage, engine displacement, SmartWay Certification, air pollution score(APS) and drive type (GHGS ~ Cmb + Displ + SW + APS + Drive). Further explanation of these variables can be found in the definition of variables. An interesting finding in searching for our model was that dependent on the mileage type, combined; highway; or city, the factors affecting the Greenhouse Gas Score differ. Our best model used the vehicle's combined mileage. Another interesting finding is the strength of the air pollution score (APS). The group assumed that APS would be one of the most significant variables in the GHGS. However, in our model, the APS was less significant than combined mileage and engine displacement.

Introduction:

With economic difficulties plaguing the automotive industry, consumers are waiting for the next wave in automotive revolutions. With gas prices averaging over $4 a gallon, alternative fuel sources are being investigated. However, with the threat of global warming escalating, a primary global concern is, are these new cars environmentally friendly?

Going green reflects a general environmental philosophy and social consciousness around saving and advancing earth's natural resources.

A carbon footprint is the measure of the impact that an activity has on the environment. It relates to the amount of greenhouse gases produced by burning fossil fuels for electricity, heating, or transportation. Carbon dioxide (CO2) gases are felt to be a primary source of climate change, with cars and trucks being responsible for 30 percent of these emissions. There are many driver options for easing into a green lifestyle, and as many states adopt stricter emissions laws, people are looking at their vehicle choice. Hybrid cars offer better mileage, fewer emissions, and savings on gas, but they can also sacrifice some power. Capable of achieving over 40 miles per gallon hybrid vehicles reduce America's dependence on

foreign oil and minimize emissions from greenhouse gases.

The aim of this project is to evaluate the extent of effectiveness of certain categorical and numerical values in determining the Green House Gas score of a vehicle. The Green House Gas score is a measure of the environmental friendliness of a car or truck based off its emission levels and fuel economy levels. With a score ranging from 1-10, with 10 being the best, these scores could be used to compare vehicles of all makes and models. Using variables such as the vehicle class, transmission, and drive (2 wheel drive vs. 4 wheel drive), various tests and analysis were run to ultimately determine whether or not each variable is efficient in the calculation of this Green House Score.

Methodology & Analysis:

Definition of Variables

The following table is a list of the variables we chose to use from the Green Vehicle Guide Dataset collected by the EPA (see reference 1 for details)

Name	Type of Variable	Definition
GHGS	Numeric	Green House Gas Score (See Description Below)
Displ	Numeric	Engine Displacement (measured in liters)
Cyl	Categorical	Number of Cylinders in the vehicles engine
Trans	Categorical	Type of Transmission in the vehicle

Drive	Categorical	2 Wheel Drive or 4 Wheel Drive
Fuel	Categorical	The type of fuel a vehicle takes
VC	Categorical	Vehicle Class (small car, suv, etc.)
APS	Numeric	Air Pollution Score (See Description Below)
City	Numeric	Miles Per Gallon in a city environment
Hwy	Numeric	Miles Per Gallon in a highway environment
Cmb	Numeric	Miles Per Gallon combined (computed by Averaging city and hwy)
SW	Categorical	Smart Way Approved (See Description Below)

Explained below are the important variables used in our model:

Green House Gas Score:

The Greenhouse Gas Score reflects a vehicle's tailpipe greenhouse gas emissions. A vehicle's CO2 emissions are based on the carbon content of the fuel used and the fuel economy of your engine. In addition to CO2, the GHG score includes the tailpipe greenhouse gas emissions of methane (CH4) and nitrous oxide (N2O), which are largely dependent on a vehicle's emission control technology and the miles traveled. The Greenhouse Gas Score ranges from 0 to 10, with a score of 0 being most harmful to the environment through greenhouse gas emissions, and 10 meaning no greenhouse gas is emitted.[1]

Air Pollution Score:

This score reflects vehicle tailpipe emissions that contribute to local and regional air pollution, creating problems such as smog, haze, and health issues. Vehicles that score a 10 are the cleanest, emitting no pollutants. Vehicles that score a 0 greatly emit harmful pollutants. (Scoring standards can be found in the appendix)[1]

Smart Way Approved:

SmartWay designation is a status earned by vehicles that have combined Air Pollution and Greenhouse Gas Scores that place them in the top tier of environmental performers. The variable is a categorical variable, stating whether a vehicle is SmartWay approved or is not.

[1] http://www.epa.gov/greenvehicles/Aboutratings.do

Engine Displacement:

Engine displacement is a measure of the volume in an internal combustion engine. Though not directly proportional to total power produced, it typically correlates strongly with output power. High engine displacement will generally result in low-fuel economy, and various governments have used the figure as a basis for taxation.[2]

Vehicle's Drive Type:

Vehicle drive type describes whether a vehicle is Four-Wheel Drive (4WD) or Two-Wheel Drive (2WD). The group's initial assumption is that a 4WD model vehicle uses more gas, therefore emits more harmful pollutants.

Choosing Our Model

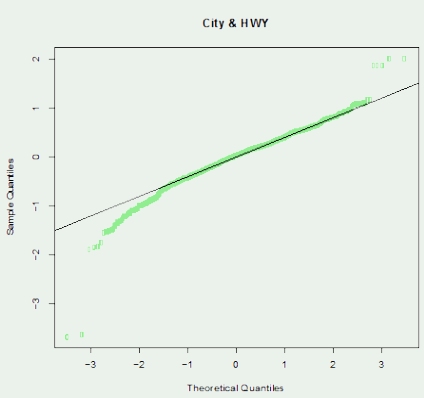
The first step in our analysis was running a series of several tests to refine our dataset. A combination of stepwise regressions and ANOVA tests helped us find a dataset that had the least variables with the highest R^2 value.

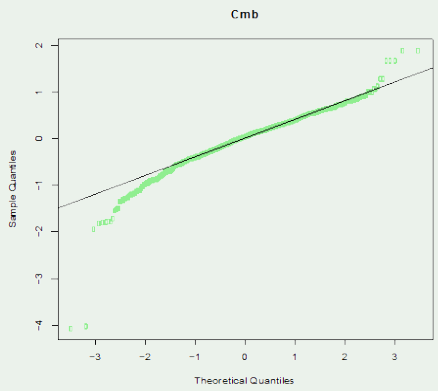
MPG Choice:

We created two different models. The two can be seen in the table below.

HWY & City Model	Cmb Model

[2] http://www.wisegeek.com/what-is-engine-displacement.htm



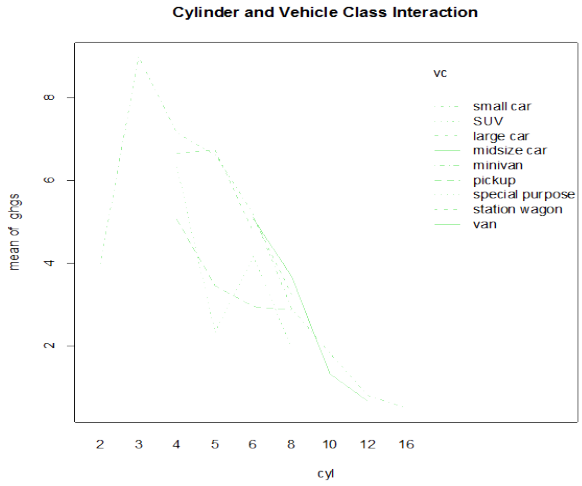


$R^2 = 0.9475$	$R^2 = 0.9488$
(See Code Appendix 1 for source)	*(See Code Appendix 2 for source)*

Although a menial difference the model with *Cmb* proved to be a better model with a higher R^2 value.

Variable Interaction:

To gain a stronger awareness of the independence of each variable we ran an amount of interaction plots. The figure below shows the most significant interaction:



The Best Model:

The goal of our project is to find which variables provide the most influence on the *GHGS.* It is clear that all of these values do have an influence because the EPA included them in the table. In the next section we began to slowly diminish the amount of variables in our

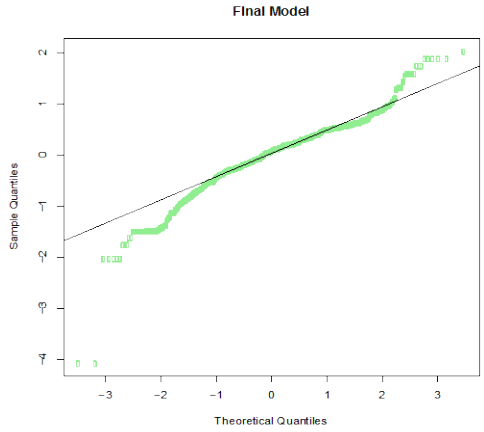
model. We did this by doing several stepwise regressions. After each test we would attempt to remove the least significant variables. If the effect on the R^2 was not significantly lower, we would keep the new model.

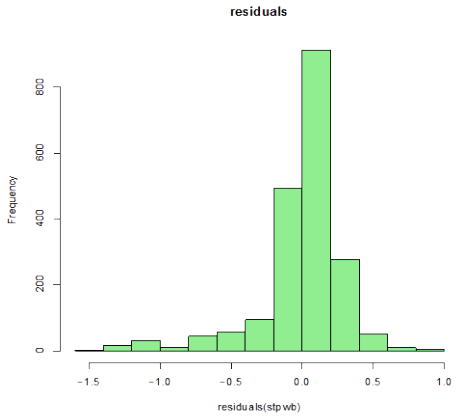
R^2 value of each new model

Origional Model	Removing trans	Removing fuel	Removing cyl	Removing vc
R^2 = .9488	R^2 = 0.947	R^2 = 0.9407	R^2= 0.9278	R^2 = 0.9265

Our final model had the following variables: *displ, drive, aps, cmb, and sw.*

This is a qqnorm plot and a histogram of the residuals:





(see code appendix 3 for source)

ANOVA

So far we had cut down our variables by almost half, and only lost less than .03 in our model's R^2 value. Our next step was to run an ANOVA test on our final model to see two things. 1 if our model was strong and 2 if there were any more variables that could be dropped out.

The results showed that all values were significant. Below is a table of the F and P values of each Variable.

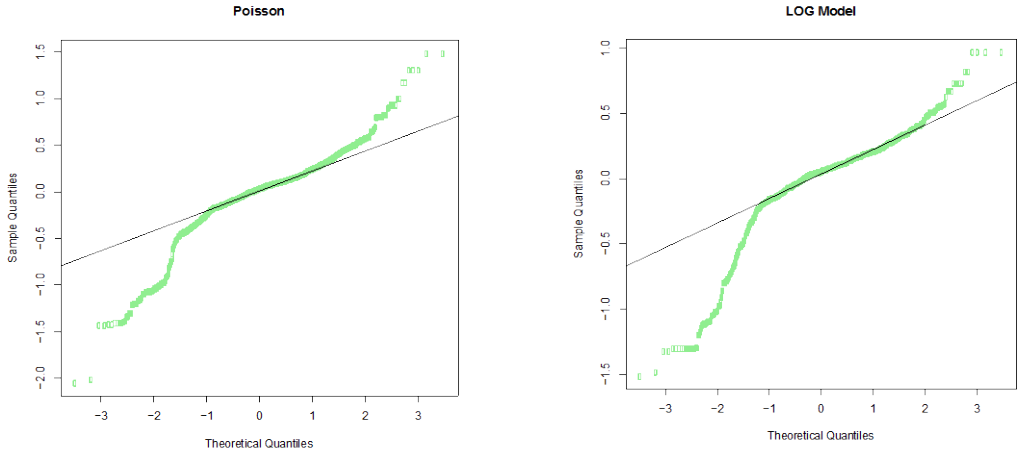
Variable Name	F Value	P Value
Cmb	5264.0145	< 2.2e-16***
Displ	510.7736	< 2.2e-16***
Sw	67.6027	3.553e-16**
Aps	8.2832	0.004044**

Drive	8.9951	0.002740**

(See Code Appendix 4 for source)

Poisson & Log

It seemed as if we had our final model. In our last step we attempted to change the family of linear regression used to Poisson. These results yielded a worse model. A similar thing happened when we tried to increment our 0 values by a small amount and take the log of them.



(See Code Appendix 5 for source)

Predicting with Our Model

We next attempted to predict the GHGS of a car by using our model. Below are the actual GHGS scores with our predicted scores.

Name	Actual	Prediction
Toyota 4Runner	4	4.10346
Ford Escape	9	8.713507
Dodge Ram	2	2.11677
Theoretical Car 1		2.11677
Theoretical Car 2		6.954193

Comparing Cmb To Hwy and City

The final thing we did was comparing stepwise regressions of our initial Cmb model to two other models with just Hwy and City. The idea behind this was to see if different values were significant when just Hwy or just City were used. The table below contains the

significant variables for each regression. The stars represent how significant the variable is.

	Cmb	Hwy	City
Aps	**		**
Cyl12			***
Cyl4	**	*	***
Cyl5	**	*	***
Cyl6	*		***
Cyl10	**	**	
Cyl12	***	***	
Displ	*	***	***
Drive4wd	***	*	***
Fueldiesel	***	***	***
fuelEthanol/Gas			**
fuelGasoline	***	*	***
Swyes	***	***	***
transAuto-6		**	
transCVT		***	
transMan-5		*	**
transMan-7	**	***	
transOther-1		***	
transSemiAuto-4		***	
transSemiAuto-6	*		**
transSemiAuto-8	***	***	***
Vclargecar	***		***
Vcmidsize	***	***	***
Vcpickup	***		***
Vcsmall car	***	***	***
Vcstation wagon	***		***
Vcvan	**		**

(See Code Appendix 7 for Source)

Results:

When paired with each other against Greenhouse Gas Score means, vehicle class and cylinder count were found to have interaction. This was the strongest, most significant interaction discovered from the data. These variables proved insignificant in the combined mileage model; however, the pair proved significant in the stepwise-created models using city and highway mileage respectably.

Our final model for predicting Greenhouse Gas Score used variables: combined mileage, engine displacement, SmartWay certification, and the vehicle's drive type. A resulting adjusted R-squared of 0.92 indicates that 92% of the variability in the Greenhouse Gas Score variable can be explained by the explanatory variables in the model. An interesting finding in the results of our tests is the level of significance of a vehicle's air pollution score as compared to other factors. We believed that the air pollution score of a vehicle would be equal, if not greater, in significance to its mileage. In the highway mileage model, the APS is one of the least significant variables.

Conclusions:

Based on the results received from the analysis, it was possible to answer the question raised in the beginning of our study; using the variables combined mileage, engine displacement, air pollution score, drive type and SmartWay certification, we were able predict, with about 90% accuracy, a vehicle's greenhouse gas score's using our final model. Our best model was derived by doing several stepwise regressions.

The model has a high R-squared value, and roughly normal residuals. Applying log transformations and using the poisson family were techniques used to try and normalize residuals, however it resulted in further deviation of the residuals. Box-Cox transformation found the lambda value to be very close to one, so further attempts at transformation were not needed.

Limitations:

The collected data set contained experiment data for over two thousand different vehicle models. Some models were missing data for mileage fields, the air pollution score field, or the greenhouse gas score. All three of these fields are important to our study, so these elements had to be removed from our data.

The provided data set also contains data on different fuel types. However, because of the recent innovation of dual fuel type vehicles, some of the models having gasoline and ethanol combined fuel types did not have sufficient data to be included in the model. Gasoline dominated fuel type, so this variable was not declared explanatory.

The EPA data set includes an "Air Pollution Score" value, but not the measure of the actual emissions of a vehicle. The vehicle's true emission score may give us a better understanding of a vehicle's green house gas score and improve our model.

References:

The vehicle data set was provided by the United States Environmental Protection Agency, an agency of the federal government of the United States charged with protecting human health and the environment, by writing and enforcing regulations based on law passed by Congress.

Code Appendix:

All code in this section uses the following R code to load the table. The table set1.csv is available at http://www.data.gov/raw/2004

```
> library(gdata)

> bw = read.csv('set1.csv')

> model = bw$Model

> displ = bw$Displ

> bw$Cyl = factor(bw$Cyl, labels = c("2", "3", "4", "5", "6", "8", "10", "12", "16"))

> cyl = bw$Cyl

> is.factor(cyl)

[1] TRUE

> trans = bw$Trans

> drive = bw$Drive

> fuel = bw$Fuel

> sa = bw$SalesArea

> stnd = bw$Stnd

> stndd = bw$StndDescription

> uid = bw$UnderhoodId

> vc = bw$VC

> aps = bw$APS

> city = bw$City

> hwy = bw$Hwy

> cmb = bw$Cmb

> ghgs = bw$GHGS

> sw = bw$SW
```

Code Appendix 1:

```
null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+city+hwy+ghgs+sw)

full.model.formula= ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+city+hwy+ghgs+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

qqnorm(residuals(stpwb), main="City & HWY", col="lightgreen")

qqline(residuals(stpwb), add=TRUE)

summary(stpwb)
```

Code Appendix 2:

```
null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+cmb+ghgs+sw)

full.model.formula= ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+cmb+ghgs+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

qqnorm(residuals(stpwb), main="Cmb", col="lightgreen")

qqline(residuals(stpwb), add=TRUE)

summary(stpwb)
```

Code Appendix 3

```
null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+cmb+sw)

full.model.formula= ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+cmb+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)


#Removing trans

null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+cyl+drive+fuel+vc+aps+cmb+sw)

full.model.formula= ghgs ~ displ+cyl+drive+fuel+vc+aps+cmb+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)


#remonving fuel

null.model= lm(ghgs ~ 1, data=bw)
```

```
full.model= lm ( ghgs ~ displ+cyl+drive+vc+aps+cmb+sw)

full.model.formula= ghgs ~ displ+cyl+drive+vc+aps+cmb+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)


#removing cyl

null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+drive+vc+aps+cmb+sw)

full.model.formula= ghgs ~ displ+drive+vc+aps+cmb+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)


#removing vc

null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+drive+aps+cmb+sw)

full.model.formula= ghgs ~ displ+drive+aps+cmb+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)

qqnorm(residuals(stpwb), main="Final Model", col="lightgreen")

qqline(residuals(stpwb), add=TRUE)

ANOVA(stpwb)
```

Code Appendix 4

```
null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+drive+aps+cmb+sw)

full.model.formula= ghgs ~ displ+drive+aps+cmb+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)

qqnorm(residuals(stpwb), main="Final Model", col="lightgreen")

qqline(residuals(stpwb), add=TRUE)

ANOVA(stpwb)
```

Code Appendix 5

```
fjfjf=glm(ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+cmb+sw, family=poisson)

qqnorm(residuals(fjfjf), main="Poisson", col="lightgreen")

qqline(residuals(fjfjf), add=TRUE)


cmbLog = cmb+.01

cmbLog = log(cmbLog)

apsLog = aps+.01

apsLog = aps = log(apsLog)

displLog = displ +.01

displLog = log(displLog)

ghgsLog = ghgs+.01

ghgsLog = log(ghgsLog)

null.model= lm(ghgsLog ~ 1)

full.model= lm (formula = ghgsLog ~ cmb + displ + sw + drive + aps)

full.model.formula= ghgs ~ displ+drive+aps+cmb+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)

qqnorm(residuals(stpwb), main="LOG Model", col="lightgreen")

qqline(residuals(stpwb), add=TRUE)

ANOVA(stpwb)
```

Code Appendix 6

```
OurModel = lm(GHGS ~ Cmb + Displ + SW + Drive + APS, data= bw)

Toy4Runner = data.frame(Cmb=c(19), Displ=c(4), SW=c("no"), Drive=c("2WD"), APS=c(6))

predict(OurModel,Toy4Runner)

# Toyota 4Runner actual GHGS = 4

#> predict(OurModel,Toy4Runner)

# 4.103406
```

```
FordEscape = data.frame(Cmb=c(32), Displ=c(2.5), SW=c("yes"), Drive=c("2WD"), APS=c(8))

predict(OurModel,FordEscape)

#Ford Escape Hybrid actual GHGS = 9

#.> predict(OurModel,FordEscape)

# 8.713507

DodgeRam = data.frame(Cmb=c(15), Displ=c(5.7), SW=c("no"), Drive=c("4WD"), APS=c(7))

predict(OurModel,DodgeRam)

#Dodge Ram1500 acutal GHGS = 2

#.> predict(OurModel,DodgeRam)

#2.11677


TheoreticalCar = data.frame(Cmb=c(25), Displ=c(2), SW=c("yes"), Drive=c("2WD"), APS=c(9))

predict(OurModel, TheoreticalCar)

# Theoretical car has significantly less MPG than the 4Runner but is similar in other aspects.

#> predict(OurModel, TheoreticalCar)

#6.954193

# Result: Higher end GHGS score, but lower than that of the 4Runner benchmark


TheoreticalCar2 = data.frame(Cmb=c(13), Displ=c(6.9), SW=c("no"), Drive=c("4WD"), APS=c(3))

predict(OurModel, TheoreticalCar2)

# Theoretical car resembles a "gas guzzling" SUV, low MPG, 4wd, high pollutants

#> predict(OurModel, TheoreticalCar2)

# 1.045673
```

Code Appendix 7

```
null.model= lm(ghgs ~ 1, data=bw)

full.model= lm ( ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+city+sw)

full.model.formula= ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+city+sw

stpwb =step( null.model, full.model.formula, direction="forward", trace=0)

summary(stpwb)

qqnorm(residuals(stpwb), main="Final Model", col="lightgreen")

qqline(residuals(stpwb), add=TRUE)

ANOVA(stpwb)
```

```
null.model= lm(ghgs ~ 1, data=bw)
full.model= lm ( ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+hwy+sw)
full.model.formula= ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+hwy+sw
stpwb =step( null.model, full.model.formula, direction="forward", trace=0)
summary(stpwb)
qqnorm(residuals(stpwb), main="Final Model", col="lightgreen")
qqline(residuals(stpwb), add=TRUE)
ANOVA(stpwb)
null.model= lm(ghgs ~ 1, data=bw)
full.model= lm ( ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+cmb+ghgs+sw)
full.model.formula= ghgs ~ displ+cyl+trans+drive+fuel+vc+aps+cmb+ghgs+sw
stpwb =step( null.model, full.model.formula, direction="forward", trace=0)
qqnorm(residuals(stpwb), main="Cmb", col="lightgreen")
qqline(residuals(stpwb), add=TRUE)
summary(stpwb)
```

Vehicle Emission Standards and Air Pollution Score

US EPA Federal Tier 2 Emission Standard Bins and California and Northeast States LEV II Emission Standards							Air Pollution Score
Standard	Vehicles	Emission Limits at Full Useful Life (100,000-120,000 miles) Maximum Allowed Grams per Mile					
		NOx	NMOG	CO	PM	HCHO	
Bin 1	LDV, LLDT, HLDT, MDPV	0.00	0.000	0.0	0.0	0.0	10
ZEV	LDV, LDET	0.00	0.000	0.0	0.0	0.0	
PZEV	LDV, LDT	0.02	0.010	1.0	0.01	0.004	9.5
SULEV II	LDV, LDT	0.02	0.010	1.0	0.01	0.004	9
Bin 2	LDV, LLDT, HLDT, MDPV	0.02	0.010	2.1	0.01	0.004	
Bin 3	LDV, LLDT, HLDT, MDPV	0.03	0.055	2.1	0.01	0.011	8
ULEV II	LDV, LDT	0.07	0.055	2.1	0.01	0.011	7
Bin 4	LDV, LLDT, HLDT, MDPV	0.04	0.070	2.1	0.01	0.011	
Bin 5	LDV, LLDT, HLDT, MDPV	0.07	0.090	4.2	0.01	0.018	6
LEV II	LDV, LDT	0.07	0.090	4.2	0.01	0.018	
Bin 6	LDV, LLDT, HLDT, MDPV	0.10	0.090	4.2	0.01	0.018	5
LEV II option 1	LDV, LDT	0.10	0.090	4.2	0.01	0.018	
SULEV II	MDV4	0.10	0.100	3.2	0.06	0.008	4
Bin 7	LDV, LLDT, HLDT, MDPV	0.15	0.090	4.2	0.02	0.018	
SULEV II	MDV5	0.20	0.117	3.7	0.06		3
Bin 8a	LDV, LLDT, HLDT, MDPV	0.20	0.125	4.2	0.02	0.018	
ULEV II	MDV4	0.20	0.143	6.4	0.06	0.016	
Bin 8b	HLDT, MDPV	0.20	0.156	4.2	0.02	0.018	
LEV II	MDV4	0.20	0.195	6.4	0.12	0.032	2
Bin 9a	LDV, LLDT	0.30	0.090	4.2	0.06	0.018	
Bin 9b	LDT2	0.30	0.130	4.2	0.06	0.018	
Bin 9c	HLDT, MDPV	0.30	0.180	4.2	0.06	0.018	
ULEV II	MDV5	0.40	0.167	7.3	0.06		
Bin 10a	LDV, LLDT	0.60	0.156	4.2	0.08	0.018	1
LEV II	MDV5	0.40	0.230	7.3	0.12		
Bin 11	MDPV	0.90	0.280	7.3	0.12	0.032	0

See Glossary in Summary of Current and Historical Emission Standards for explanation of terms.