Energy Usage in 1978 Automobiles

This iteration is for LaTeX and uses code fences aka strict markdown. There are gains and losses:

Gains

- LaTeX looks nice

Losses

- LaTeX is difficult to use
- the document gets harder to read
- the document no mixes markdown, LaTeX, and Stata

This is a meant to be a very simple exposition about modeling energy usage using Stata's auto dataset. To make things work more nicely, let's pretend that this is some sort of sample of measurements, so that when we talk about "average energy consumption", it will make some sense.

Let's open the **auto** dataset, and look at its structure.

```
. sysuse auto, clear
(1978 Automobile Data)

. describe

Contains data from /Applications/AAApplications/MathTools/Stata15/ado/base/a/au
> to.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          13 Apr 2016 17:45
 size:         3,182                          (_dta has notes)
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
make            str18   %-18s                 Make and Model
price           int     %8.0gc                Price
mpg             int     %8.0g                 Mileage (mpg)
rep78           int     %8.0g                 Repair Record 1978
headroom        float   %6.1f                 Headroom (in.)
trunk           int     %8.0g                 Trunk space (cu. ft.)
weight          int     %8.0gc                Weight (lbs.)
length          int     %8.0g                 Length (in.)
turn            int     %8.0g                 Turn Circle (ft.)
displacement    int     %8.0g                 Displacement (cu. in.)
gear_ratio      float   %6.2f                 Gear Ratio
foreign         byte    %8.0g      origin     Car type
-------------------------------------------------------------------------------
Sorted by: foreign
```

We could use a `codebook` command here to look at all the variables, but it will take up too much space. Let's do this instead:

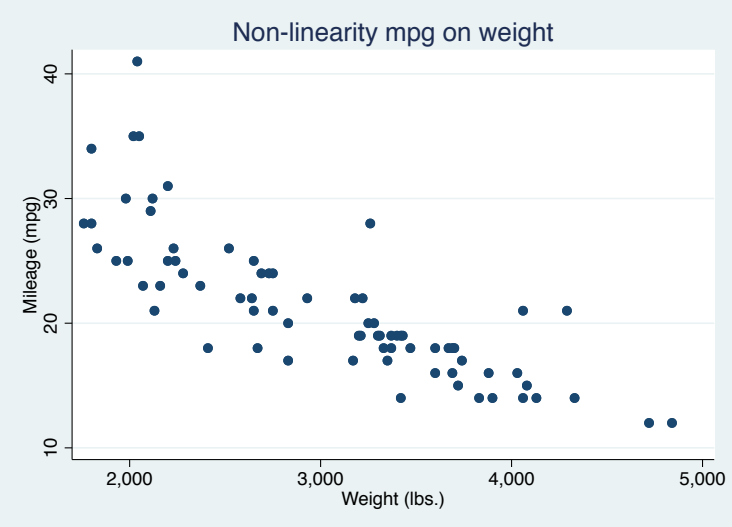
```
. codebook, compact
Variable      Obs Unique      Mean   Min    Max  Label
-------------------------------------------------------------------------------
_____________________________________________________________
```

make	74	74	.	.	.	Make and Model
price	74	74	6165.257	3291	15906	Price
mpg	74	21	21.2973	12	41	Mileage (mpg)
rep78	69	5	3.405797	1	5	Repair Record 1978
headroom	74	8	2.993243	1.5	5	Headroom (in.)
trunk	74	18	13.75676	5	23	Trunk space (cu. ft.)
weight	74	64	3019.459	1760	4840	Weight (lbs.)
length	74	47	187.9324	142	233	Length (in.)
turn	74	18	39.64865	31	51	Turn Circle (ft.)
displacement	74	31	197.2973	79	425	Displacement (cu. in.)
gear_ratio	74	36	3.014865	2.19	3.89	Gear Ratio
foreign	74	2	.2972973	0	1	Car type

For those unfamiliar with the system of weights and measures used in the United Stated (and Liberia), the important conversions to remember are that

- 1 mile = 1605 meters = 1.605 km
- 1 pound = 453.6 grams = 0.4536 kg
- 1 gallon = 3.7854 liters or about 15/4 liters
- 1 inch = 2.54 cm
- Hence 1 cubic inch = 16.387064 cc or about 16 3/8 cc
- 1 foot = 12 inches = 30.48cm
- Hence 1 cubic foot = 28316.847 cc or about 28 1/3 liters

One other oddity in the so-called traditional (or Standard or English or Imperial) system, is that energy usage is measured in miles per gallon (mpg). This is not good for analysis, because it makes for a non-linear relationship between weight and energy. This can be seen in following graph:



To make the analysis work better, we should make a variable measuring gallons use per 100 miles driven:

```
. gen gp100m = 100/mpg, before(mpg)
. label variable gp100m "Gallons per 100 miles"
```

One last conversion 1 gallon per 100 miles is about 75/32 (= 2.344) liters per 100 km.

Let's take a look at various variables by whether the cars are from the US (*domestic*), or whether they are from outside the US (*foreign*).

```
. tabstat gp100m weight length turn displacement gear_ratio, ///
>   statistics( mean sd count ) by(foreign)
Summary statistics: mean, sd, N
  by categories of: foreign (Car type)
```

foreign	gp100m	weight	length	turn	displa~t	gear_r~o
Domestic	5.318155	3317.115	196.1346	41.44231	233.7115	2.806538
	1.224346	695.3637	20.04605	3.967582	85.26299	.3359556
	52	52	52	52	52	52
Foreign	4.312848	2315.909	168.5455	35.40909	111.2273	3.507273
	1.144388	433.0035	13.68255	1.501082	24.88054	.2969076
	22	22	22	22	22	22
Total	5.01928	3019.459	187.9324	39.64865	197.2973	3.014865
	1.279856	777.1936	22.26634	4.399354	91.83722	.4562871
	74	74	74	74	74	74

This works, but it would be nice to have a table which makes it easier to see comparisons. For a simple example (with fewer statistics), we can use Ian Watson's `tabout`, version 3, from http://tabout.net.au. To facilitate the options needed for rerunning the command for different output types, the options for generating the command have been put in the file `tabout_oneway.options`.

Mean values for US and Non-US cars

	Gp100M	Weight	Displacement	Gear Ratio
Car type				
Domestic (70%)	5.32	3,317.1	233.7	2.807
Foreign (29%)	4.31	2,315.9	111.2	3.507
Total (100%)	5.02	3,019.5	197.3	3.015

Source: auto.dta

If time permits, we should be able to make a more-complete version of this table.

Before modelling, we should take a look to see if there could be collinearities in the predictors.

Finally, how about modelling let's first run a regression with many variables and then store the results

```
. regress gp100m weight displacement gear_ratio foreign
```

Source	SS	df	MS	Number of obs	=	74
				F(4, 69)	=	56.84
Model	91.7374232	4	22.9343558	Prob > F	=	0.0000
Residual	27.8388375	69	.403461414	R-squared	=	0.7672
				Adj R-squared	=	0.7537
Total	119.576261	73	1.63803097	Root MSE	=	.63519

gp100m	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	.0014428	.000216	6.68	0.000	.0010118	.0018737
displacement	.0012388	.0021161	0.59	0.560	-.0029828	.0054603
gear_ratio	-.2037991	.3258603	-0.63	0.534	-.8538726	.4462744
foreign	.733736	.2301493	3.19	0.002	.2746007	1.192871
_cons	.8147969	1.239181	0.66	0.513	-1.657301	3.286895

We can see that, as expected, heavier cars take more energy to move. Perhaps unexpectedly, non-US cars use more gas at the same weight. It appears that we can throw out both `displacement` and `gear_ratio` as predictors and fit a simpler model.

```
. regress gp100m weight foreign
```

Source	SS	df	MS	Number of obs	=	74
				F(2, 71)	=	113.97
Model	91.1761694	2	45.5880847	Prob > F	=	0.0000
Residual	28.4000913	71	.400001287	R-squared	=	0.7625
				Adj R-squared	=	0.7558
Total	119.576261	73	1.63803097	Root MSE	=	.63246

gp100m	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
weight	.0016254	.0001183	13.74	0.000	.0013896	.0018612
foreign	.6220535	.1997381	3.11	0.003	.2237871	1.02032
_cons	-.0734839	.4019932	-0.18	0.855	-.8750354	.7280677

We can put these coefficients in a table

	(1) gp100m	(2) gp100m
weight	0.00144***	0.00163***
	(6.68)	(13.74)
displacement	0.00124	
	(0.59)	
gear_ratio	-0.204	
	(-0.63)	

foreign	0.734**	0.622**
	(3.19)	(3.11)
_cons	0.815	-0.0735
	(0.66)	(-0.18)
N	74	74

t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

From the simple model, cars from 40 years ago used 0.163 gallons per mile per extra 100 pounds, on average. Also, non-US cars use about 0.622 more gallons per mile, on average, all other things being equal.