Car Mileage Data

Arden Miller

Executive Summary

Data measured on cars (1999 models) was used to create a model that relates city mileage to price, weight, and engine displacement. This model indicates a negative relationship between city mileage and both price and engine displacement. The model indicates that very light cars (weights around 2000 pounds) have the highest mileages whereas that cars that weigh approximately 3800 pounds have the lowest mileage. Surprisingly, the model indicates that cars that weigh from 4500-5000 pound get, on average, somewhat better mileage than cars that weigh around 3800 pounds.

1999 Car Data

The data analysed for this report represents data on 138 cars that were taken from *Road and Track's "The Complete '99 Car Buyer's Guide"*. The following measurements are used:

PRICE:	price in dollars (US),
WEIGHT:	weight in pounds,
CITY:	mileage (miles per gallon) in city driving,
DISP:	displacement in cubic centimetres,
COMP:	compression ratio as value to 1,
HP:	horsepower at 6300 rpm,
TORQ:	torque at 5200 rpm,
TRANS:	transmission (1 = automatic, 0 = manual),
CYL:	number of cylinders.

This report focuses on how mileage in city driving (CITY) is related to the other variables. For this data, values of CITY range from 9 to 41 mpg with a mean of 23.4 mpg. A histogram of CITY as well as histograms of the other measurements are given in Figure 1. The histogram for CITY indicates that well over half the cars got between 15 and 25 miles per gallon for city driving and that the distribution is right skewed. The other histograms indicate that PRICE, DISP, HP, TORQ, and CYL all have right skewed distributions as well whereas the distributions of WEIGHT and COMP are reasonably symmetric. Note that TRANS is a indicator variable with 0 representing manual and 1 representing automatic. Approximately 55% of the cars in this data set had manual transmissions.

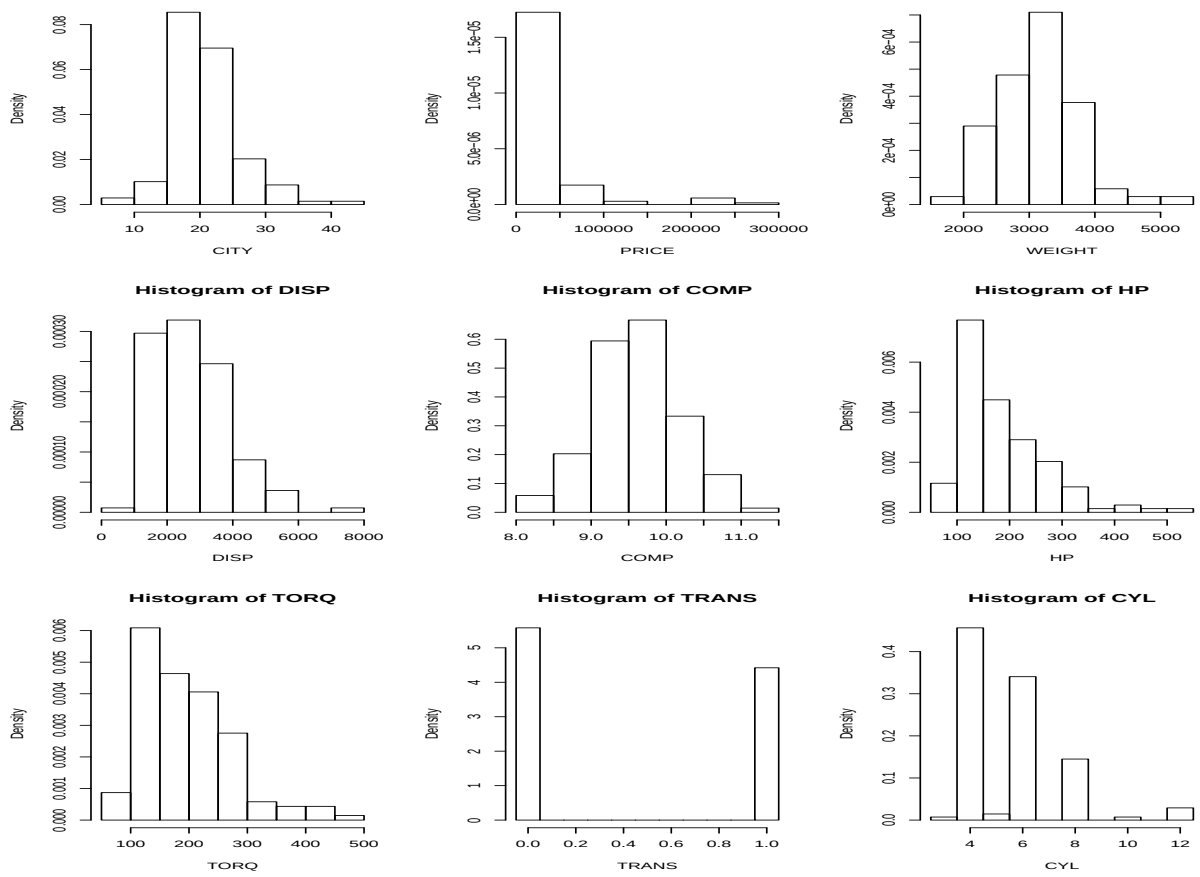


Figure 1: Histograms of Car Data.

Figure 2 contains pairwise scatter-plots of these measurements. This plot indicates that there is a clear relationship between CITY and each of the other variables except compression ratio (COMP). CITY is negatively related to each of PRICE, WEIGHT, DISP, HP, TORQ, and CYL. In most cases these relationships are curved rather than linear. The plot of CITY versus TRANS indicates mileage is generally higher for manual transmissions. The other pairwise scatterplots (those that do not involve CITY) show that there are strong relationships between many pairs of the variables.

1 Modeling the Car Mileage

Two regression models that relate CITY to the other variables were identified. These models should only be used to make predictions for cars that have similar characteristics to the cars included to the data set. For example, all the cars in the data set were 1999 models. Thus strictly speaking these regression models are only valid for cars from 1999.

The first model is:

$$\begin{aligned}\log(\mathsf{CITY}) &= 4.92 - 2.40 \times 10^{-6}(\mathsf{PRICE}) - 8.91 \times 10^{-4}(\mathsf{WEIGHT}) + 1.19 \times 10^{-7}(\mathsf{WEIGHT}^2) \\ &\quad -4.23 \times 10^{-5}(\mathsf{DISP}) + 8.52 \times 10^{-4}(\mathsf{HP}) - 1.32 \times 10^{-3}(\mathsf{TORQ})\end{aligned}$$

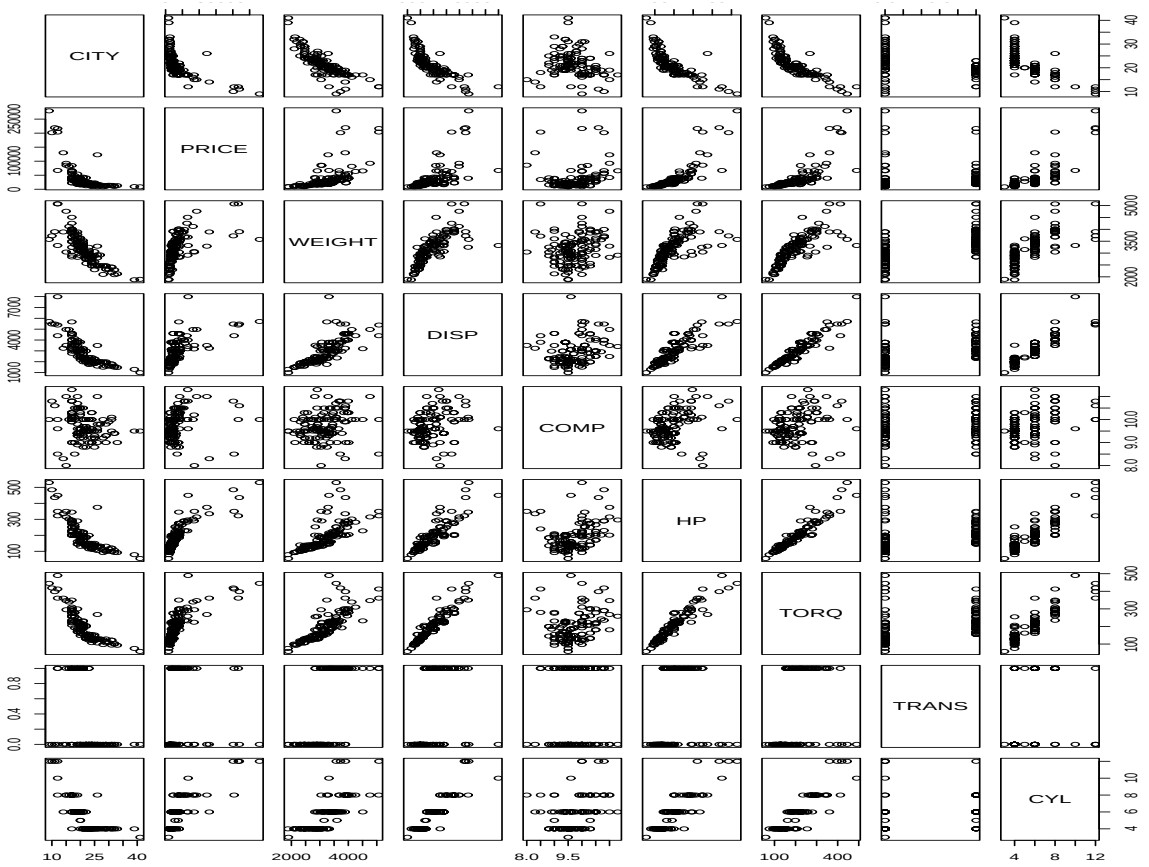


Figure 2: Pairwise Scatterplots of Car Data.

This model predicts log(CITY) using PRICE, WEIGHT, DISP, HP, and TORQ as explanatory variables. To convert the log(CITY) predictions to predictions for CITY it is necessary to apply the exponential function. Diagnostic checks indicate that this model satisfies the assumptions for a regression model reasonably well except that it is strongly influenced by a single observation (Ferrari F355 Berlinetta). This observation has a particularly big impact on the estimated coefficients for HP and TORQ. In fact if this observation is removed there is no longer evidence that these two variables are needed in the model. In this case, the model can be simplified to:

$$\begin{aligned}\log(\mathsf{CITY}) &= 4.98 - 2.57 \times 10^{-6}(\mathsf{PRICE}) - 9.20 \times 10^{-4}(\mathsf{WEIGHT}) + 1.21 \times 10^{-7}(\mathsf{WEIGHT}^2) \\ &\quad -6.69 \times 10^{-5}(\mathsf{DISP})\end{aligned}$$

I would recommend using this simpler model since it doesn't seem reasonable to include two extra variables in the model on the basis of 1 observation. To evaluate, the precision of predictions made using the above model, 95% prediction intervals were calculated for each point in the data set. The width of these intervals varied from 2.6 to 10.1 with a mean of 5.8.

Figure 3 indicates how the model relates CITY to each of the explanatory variables. For each plot, one of the explanatory variables is varied over the range of values it takes in the data set while the remaining explanatory variables are held constant at their mean values. The plots for PRICE and DISP show that the predicted mileage decreases as each of these regressors is

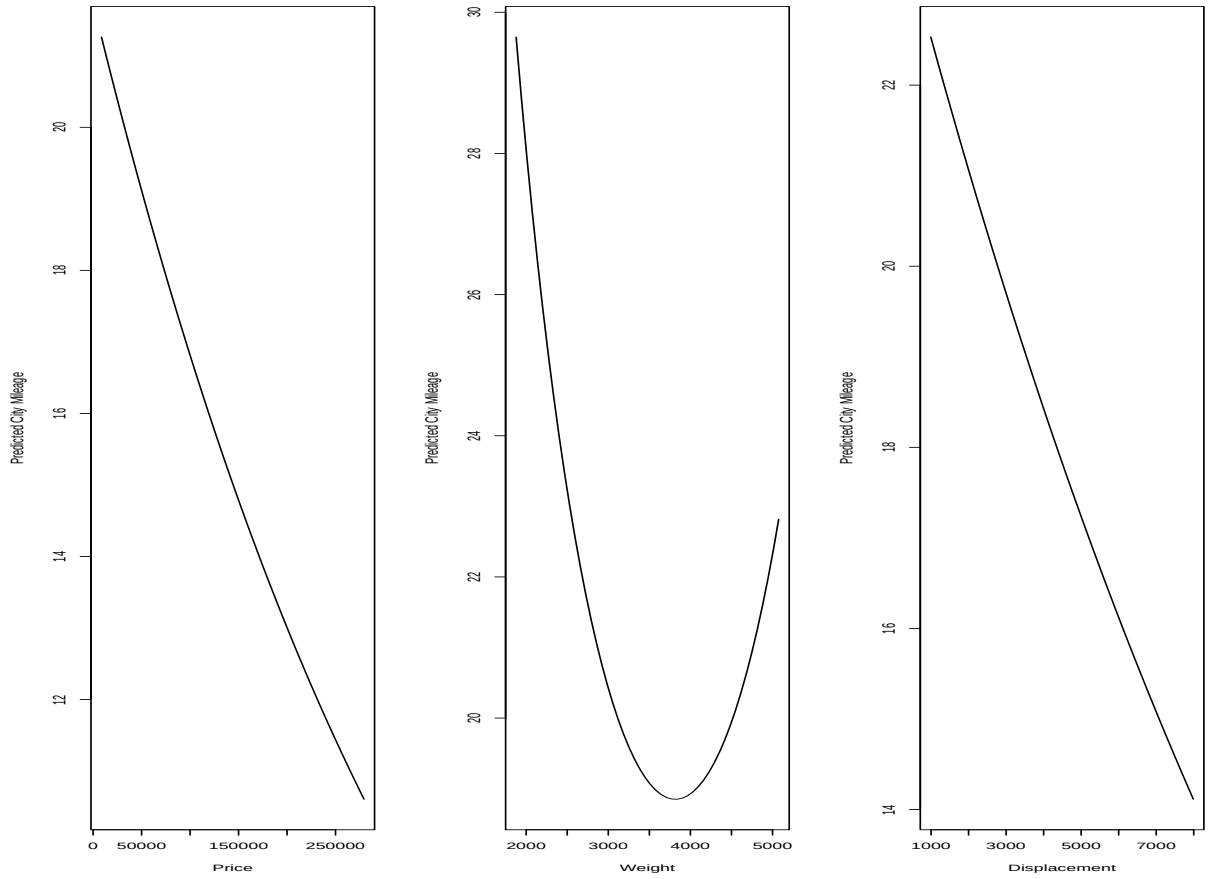


Figure 3: Plots of Predicted Mileage versus Predictors.

increased. The plot for WEIGHT indicates that predicted mileage is at minimum for cars that have WEIGHT $\approx$ 3800 pounds.

Statistical Appendix

The pairs plot for this data indicates that their are strong pairwise relationships among the explanatory. This strongly suggests that multicollinearity may be a problem with this data. This is confirmed by the VIF's:

	PRICE	WEIGHT	DISP	COMP	HP	TORQ	TRANS	CYL
VIF	6.4	5.5	20.8	1.5	23.0	46.1	2.4	11.7

These show that DISP, HP, TORQ, and CYL are severely affected (VIF's > 10) by multicollinearity and thatPRICE and WEIGHT are moderately affected (VIF's between 5 and 10). variables are clearly related to each other. As a result it is unlikely that all of the explanatory variables will be needed in the model for CITY.

I started by fitting the basic multiple regression model for CITY using linear terms for each of the explanatory variables. As expected the output for this model had non-significant P-values for most of the regressors (all except WEIGHT and HP). The plotted of residuals versus fitted values shows clear evidence of non-linearity. The output from the "funnel" command (slope= .85) indicated that a log transformation of the response may be suitable for the response.

First, I tried fitting the full model using log(CITY) as the response. This model still only has two significant coefficients (WEIGHT and HP). I tried producing partial plots which indicated That squared terms might be useful for WEIGHT, COMP, and CYL. I tried adding these to the full model and found the WEIGHT^2 was very significant but the other two terms were not. At this point, I decided to use the `all.poss.regs` command in *R* to identify the best subset models for this data - I added a column for the WEIGHT^2. The output was:

```
> all.poss.regs(X,y)
   rssp sigma2 adjRsq      Cp PRICE WEIGHT DISP COMP HP TORQ TRANS CYL WT2
1 1.451  0.011  0.809  87.572     0      0    0    0  0    1     0   0   0
1 1.937  0.014  0.745 161.647     0      0    0    0  1    0     0   0   0
1 2.116  0.016  0.721 189.103     0      0    1    0  0    0     0   0   0
2 1.277  0.009  0.830  62.988     0      1    0    0  1    0     0   0   0
2 1.284  0.010  0.829  63.990     0      1    0    0  0    1     0   0   0
2 1.359  0.010  0.819  75.538     0      0    0    0  0    1     0   0   1
3 1.110  0.008  0.851  39.430     0      1    0    0  0    1     0   0   1
3 1.149  0.009  0.846  45.466     0      1    0    0  1    0     0   0   1
3 1.175  0.009  0.843  49.345     1      1    0    0  0    0     0   0   1
4 0.883  0.007  0.881   6.762     1      1    1    0  0    0     0   0   1
4 0.899  0.007  0.879   9.223     1      1    0    0  0    1     0   0   1
4 1.034  0.008  0.861  29.877     1      1    0    0  1    0     0   0   1
5 0.869  0.007  0.882   6.652     1      1    0    0  1    1     0   0   1
5 0.870  0.007  0.882   6.772     1      1    1    0  0    1     0   0   1
5 0.871  0.007  0.882   7.033     1      1    1    0  0    0     0   1   1
6 0.842  0.006  0.885   4.594     1      1    1    0  1    1     0   0   1
6 0.863  0.007  0.882   7.671     1      1    1    1  0    1     0   0   1
6 0.863  0.007  0.882   7.803     1      1    1    0  0    1     0   1   1
7 0.840  0.006  0.884   6.230     1      1    1    0  1    1     1   0   1
7 0.840  0.006  0.884   6.287     1      1    1    0  1    1     0   1   1
7 0.842  0.006  0.884   6.579     1      1    1    1  1    1     0   0   1
8 0.838  0.006  0.883   8.000     1      1    1    0  1    1     1   1   1
8 0.840  0.007  0.883   8.213     1      1    1    1  1    1     1   0   1
8 0.840  0.007  0.883   8.287     1      1    1    1  1    1     0   1   1
9 0.838  0.007  0.883  10.000     1      1    1    1  1    1     1   1   1
```

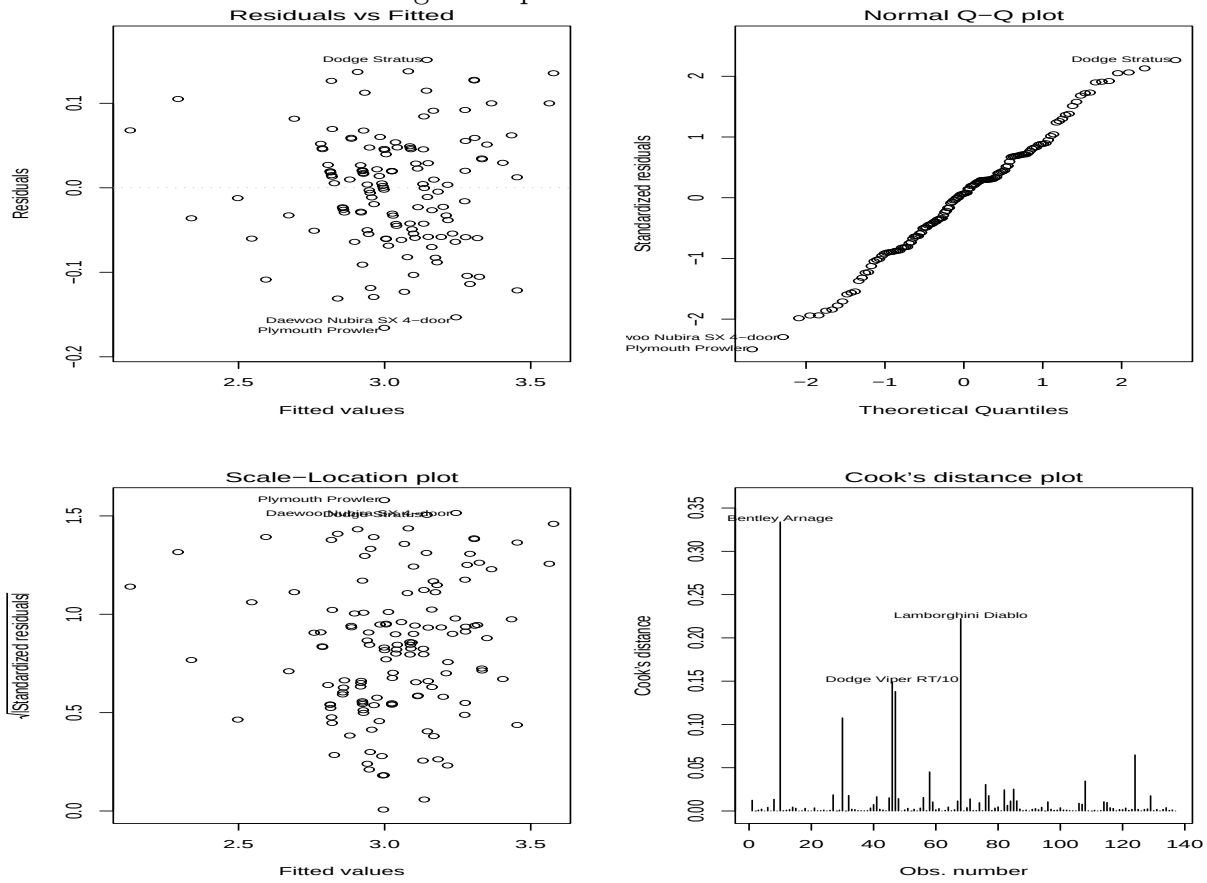
I decided to try the model that uses PRICE, WEIGHT, WEIGHT^2, DISP, HP, and TORQ since it had the minimum Cp. The output for the fitted model from *R* is:

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.924e+00  1.616e-01  30.465  < 2e-16 ***
PRICE       -2.399e-06  3.960e-07  -6.058 1.37e-08 ***
WEIGHT      -8.909e-04  1.058e-04  -8.425 5.60e-14 ***
I(WEIGHT^2)  1.193e-07  1.609e-08   7.414 1.36e-11 ***
```

```
DISP        -4.267e-05  2.099e-05  -2.033   0.0441 *
HP           8.524e-04  4.132e-04   2.063   0.0411 *
TORQ        -1.318e-03  5.382e-04  -2.450   0.0156 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08019 on 131 degrees of freedom
Multiple R-Squared: 0.8897, Adjusted R-squared: 0.8847
F-statistic: 176.2 on 6 and 131 DF,  p-value:     0
```

Diagnostic plots for this model are:



On a positive note, these plots indicate no problems with linearity, non-constant variance, or non-Normality. However they do indicate there is one observation (Ferrari F355 Berlinetta) that is an outlier and has an extremely large value of Cook's Distance. This point clearly has a very large impact on the fitted model.

If this point is removed then the coefficients for DISP and TORQ both become not significant. I dropped these variables one at a time and was left with the model that uses PRICE, WEIGHT, WEIGHT^2, and HP. I also tried using all.poss.regs and it identifies this is a sensible model.

```
Call:
lm(formula = log(CITY) ~ PRICE + WEIGHT + I(WEIGHT^2) + DISP,
    data = c99.df[-47, ])
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.978e+00  1.235e-01  40.319  < 2e-16 ***
PRICE       -2.569e-06  1.939e-07 -13.252  < 2e-16 ***
WEIGHT      -9.204e-04  7.777e-05 -11.835  < 2e-16 ***
I(WEIGHT^2)  1.206e-07  1.156e-08  10.436  < 2e-16 ***
DISP        -6.685e-05  8.792e-06  -7.604 4.79e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.0673 on 132 degrees of freedom
Multiple R-Squared: 0.9212,     Adjusted R-squared: 0.9188
F-statistic: 385.9 on 4 and 132 DF,  p-value:     0
```

Diagnostic plots for this model are: