Linear Regression Models

Statistics 220

Spring 2005

Linear Regression Models

Notation:

- y: response or outcome variable
- $x = (x_1, \ldots, x_k)$: explanatory or predictor variables. These may be continuous or discrete.

Data model: For observation $i, i = 1, \ldots, n$

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + e_i$$

Note this model doesn't explicitly include an intercept. An intercept can be included by setting $x_{i1} = 1$ for all i.

The common assumption for the error terms is

$$e_i \overset{iid}{\sim} N(0, \sigma^2)$$

This can be relaxed allowing for nonconstant variance and correlation of the errors.

Note that this model is conditioning on the x's and doesn't consider how the x's are generated. They could be random, or deterministic, as in an experiment.

Conditional Modelling

In what follows, our models will be conditional on X. When can we do this.

Assume that $X \sim p(X|\psi)$ and $y|X \sim p(y|X,\theta)$ and assume that ψ and θ have no common components.

In addition, assume that $p(\psi,\theta) = p(\psi)p(\theta)$ (i.e. ψ and θ are independent apriori).

Thus

$$\begin{aligned} p(\psi,\theta|X,y) &= \frac{p(X|\psi)p(\psi)p(y|X,\theta)p(\theta)}{p(X)p(y|X)} \\ &= p(\psi|X)p(\theta|X,y) \end{aligned}$$

which implies to learn about θ, we only need to look at $p(\theta|X,y)$.

As earlier

$$p(\theta|X, y) \propto p(y|X, \theta)p(\theta)$$

In much of what that follows, as in the text the dependency on X will be suppressed for notational clarity

Note that this situation does not include the case where X is measured with error. For example

$$\begin{aligned} y|X &\sim N(X\beta, \sigma^2 I) \\ X_i^{obs}|X_i &\overset{ind}{\sim} N(X_i, \Sigma) \\ X &\sim P(X|\psi) \end{aligned}$$

In this case, the posterior distribution of (β, σ^2) will depend on Σ.

Bayesian Analysis of Classical Regression Model

Data model:

$$y|\beta, \sigma^2, X \sim N(X\beta, \sigma^2 I)$$

where I is the $n \times n$ identity matrix

Prior:

Lets start with a convenient non-informative prior

$$p(\beta, \sigma^2|X) \propto \sigma^{-2} = \frac{1}{\sigma^2}$$

This is equivalent to saying $(\beta, \log \sigma)$ are uniform.

This prior is reasonable with n is large and k (the number of βs) is small. It may not do well with smaller sample size.

Posterior Distribution:

As with earlier normal based models, we want to take a hierarchical approach to the posterior. That is

$$p(\beta, \sigma^2|y) = p(\sigma^2|y)p(\beta|\sigma^2, y)$$

First, lets deal with $\beta|\sigma^2, y$

$$\beta|\sigma^2, y \sim N(\hat{\beta}, V_\beta\sigma^2)$$

where

$$\begin{aligned} \hat{\beta} &= (X^TX)^{-1}X^Ty \\ V_\beta &= (X^TX)^{-1} \end{aligned}$$

Next, $\sigma^2|y$ is

$$\sigma^2|y \sim \text{Inv}{-}\chi^2(n-k, s^2)$$

It can be shown that $\beta|y$ is Multivariate $t_{n-k}(\hat{\beta}, V_\beta s^2)$ (with dimension k)

Note that these results agree with the standard frequentist regression estimates.

Note that is this case the posterior distribution of (β, σ^2) is proper if

1. $n > k$ (More observations than predictor variables)

2. $\text{rank}(X) = k$ (None of the xs are a linear combination of the others.)

These are the standard conditions for existence of MLEs in standard regression models.

Sampling From the Posterior

While it is possible to sample directly from the distribution $p(\beta|y)$, which is usually the posterior distribution of interest, it is more usual to use the following algorithm

1. Sample $\sigma^{2(j)} \sim \text{Inv}{-}\chi^2(n-k, s^2)$
2. Sample $\beta^{(j)} \sim N(\hat{\beta}, V_\beta \sigma^{2(j)})$

Then $(\beta^{(j)}, \sigma^{2(j)})$ is a draw from $p(\beta, \sigma^2|y)$

Note that this is an exact analog to the sampling scheme used for the model

$$\begin{aligned} y_i|\mu, \sigma^2 &\overset{iid}{\sim} N(\mu, \sigma^2) \\ p(\mu, \sigma) &\propto \frac{1}{\sigma^2} \end{aligned}$$

To implement this, you can use the scheme discussed in the book based the QR factorization.

If you are using `R`, you can use the following code to implement sampling (β, σ^2). This assume that you have analyzed the model with the `lm()` command and stored the results in `lmout` and the library `MASS` has been loaded.

```
betahat <- coef(lmout)
df <- summary(lmout)$df
s2 <- (summary(lmout)$sigma)^2
Vbeta <- vcov(lmout)/s2

sigma2 <- 1/rgamma(n,df[2]/2, s2*df[2]/2)

# Should be a way of vectorizing this code for generating beta

beta <- matrix(0, ncol=df[1], nrow=n)
for(i in 1:n) beta[i,] <- mvrnorm(df[1], betahat, sigma2[i]*Vbeta)
```

Note that `R` implements `lm` via the QR factorization, so most of the advantages discussed in the text will occur.

Posterior Predictive Distribution

One of the advantages of using the above scheme for simulating posterior βs is that it is easy to sample from the posterior predictive distribution as

$$p(\tilde{y}|\beta, \sigma^2, y) = p(\tilde{y}|\beta, \sigma^2)$$

For example, to simulate y^{rep} for model checking, add the step

3. Sample $y^{rep(j)} \sim N(X\beta^{(j)}, V_\beta \sigma^{2(j)})$

The exact distribution of $\tilde{y}$ can be determined in this case (assume want evaluated with $\tilde{X}$).

First

$$\begin{aligned} E[\tilde{y}|\sigma^2, y] &= E[E[\tilde{y}|, \beta, \sigma^2, y]|\sigma^2, y] \\ &= E[\tilde{X}\beta|\sigma^2, y] \\ &= \tilde{X}\hat{\beta} \end{aligned}$$

Next

$$\begin{aligned} \mathrm{Var}(\tilde{y}|\sigma^2, y) &= E[\mathrm{Var}(\tilde{y}|, \beta, \sigma^2, y)|\sigma^2, y] + \mathrm{Var}(E[\tilde{y}|, \beta, \sigma^2, y]|\sigma^2, y) \\ &= E[\sigma^2 I|\sigma^2, y] + \mathrm{Var}(\tilde{X}\beta|\sigma^2, y) \\ &= (I + \tilde{X}V_\beta\tilde{X}^T)\sigma^2 \end{aligned}$$

Also note that $\tilde{y}|\sigma^2, y$ is normal with this mean and variance.

Now averaging over the posterior distribution of $\sigma^2|y$, gives that $\tilde{y}|y$ is Multivariate $t_{n-k}(\tilde{X}\hat{\beta}, (I + \tilde{X}V_\beta\tilde{X}^T)s^2)$ (with dimension $= \mathrm{rows}(\tilde{X})$).

Note that this is equivalent to what we get for prediction intervals in standard normal linear regression.

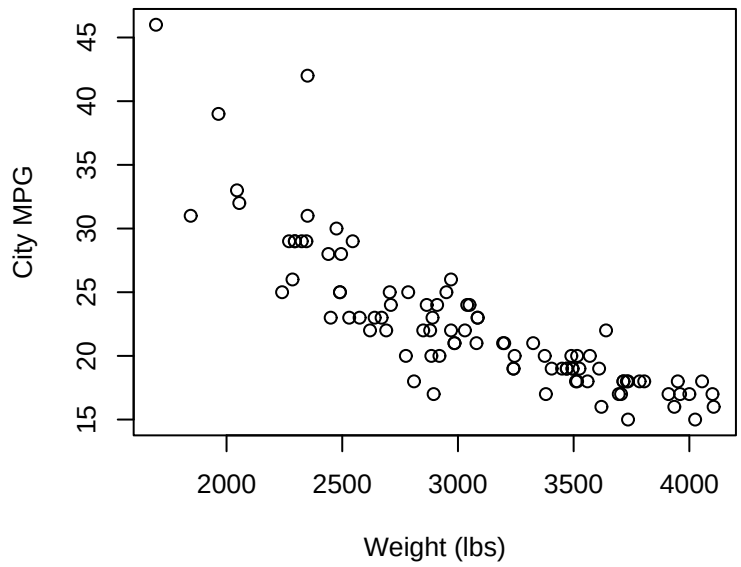
$$x^T\beta \pm t^*_{n-k}s\sqrt{1 + x^T(X^TX)^{-1}x}$$

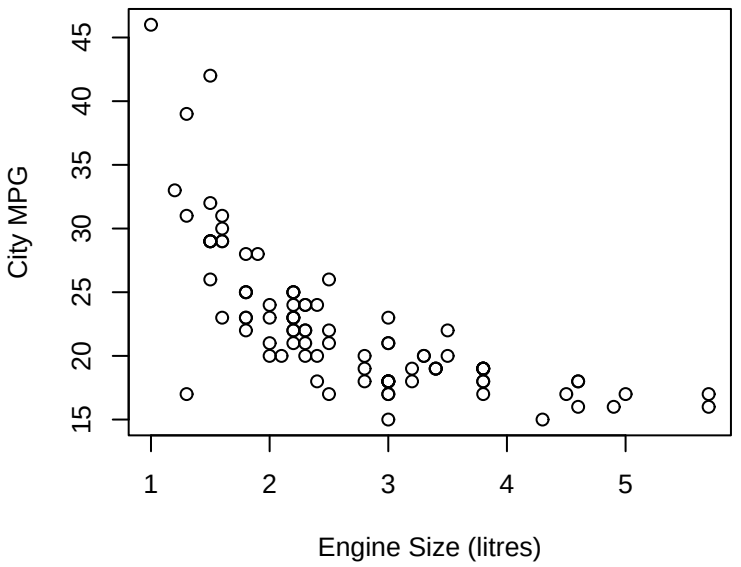
Example

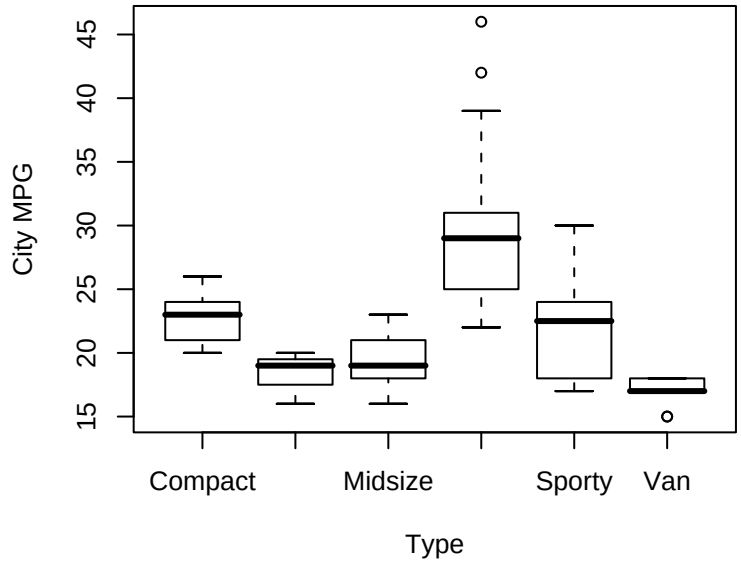
EPA Gas Ratings for 1993 Model Year

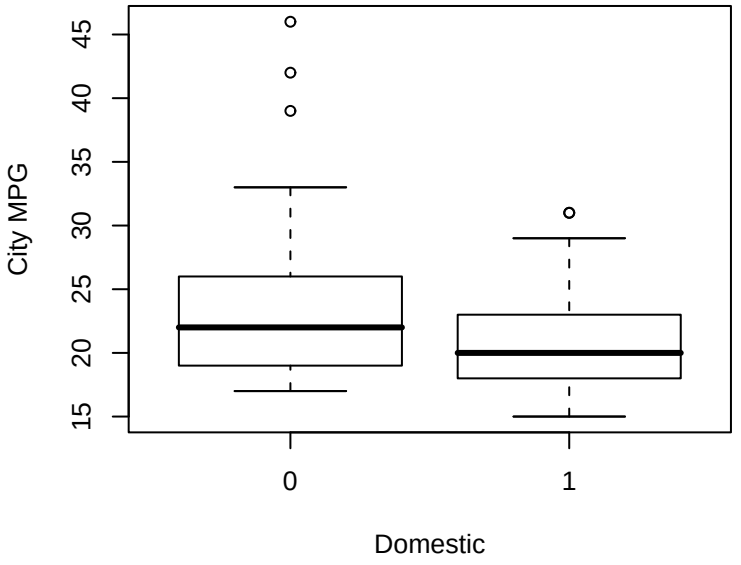
93 different car models were examined. We will focus on the EPA City MPG ratings

- $y =$ City MPG
- Weight (in lbs)
- Engine Size (in litres)
- Type: Compact, Large, Midsize, Small, Sporty
- Domestic: 0 = Foreign, 1 = Domestic


City MPG
15
20
25
30
35
40
45
2000
2500
3000
3500
4000
Weight (lbs)


City MPG
15
20
25
30
35
40
45
1
2
3
4
5
Engine Size (litres)


City MPG
15
20
25
30
35
40
45
Compact
Midsize
Sporty
Van
Type


City MPG
15
20
25
30
35
40
45
0
1
Domestic

The usual considerations for linear regression occur with a Bayesian approach. These include

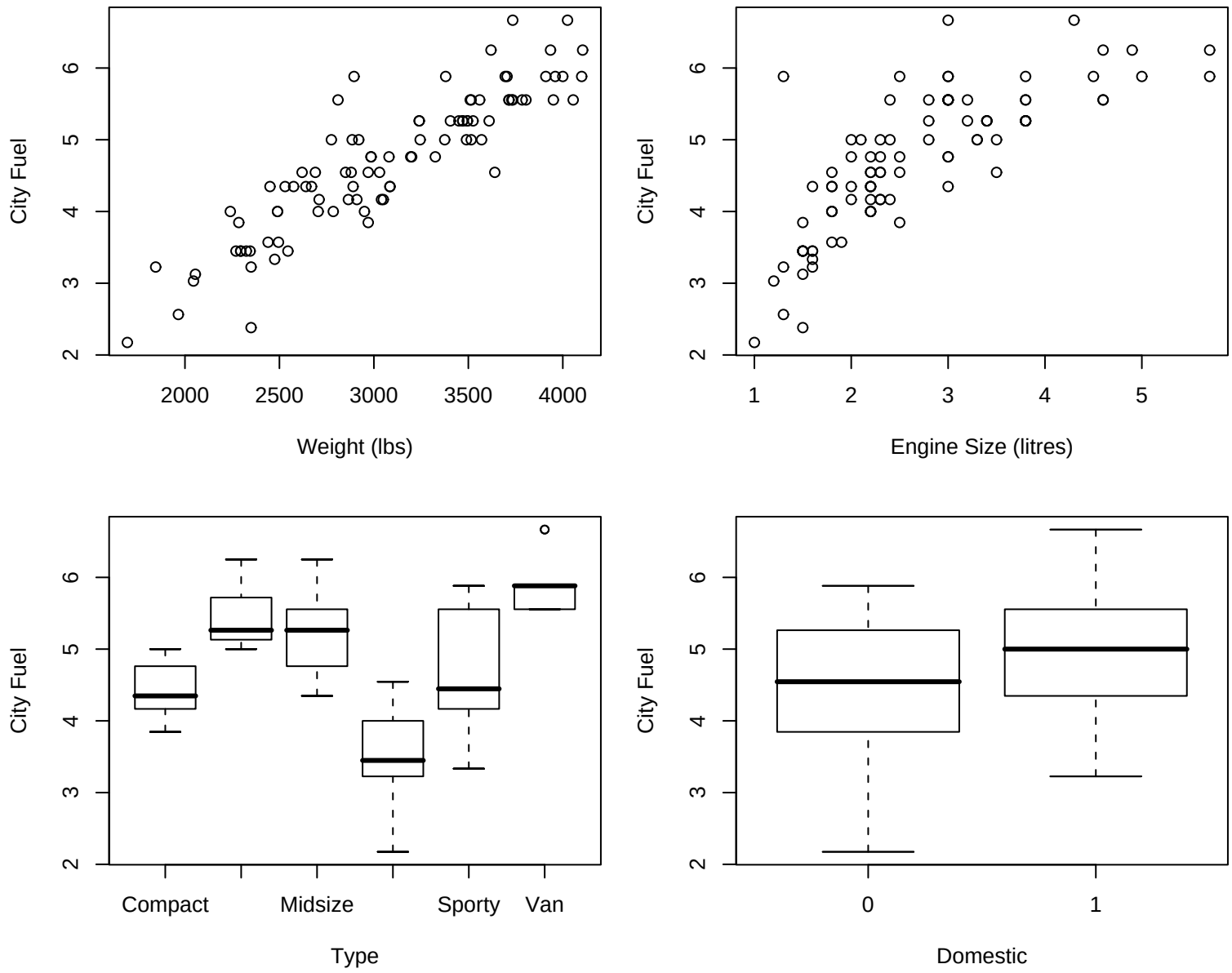
- Linearity
- Interactions
- Indicator variables for categorical predictors
- Which variables to include in model
- Distributional assumptions (e.g. normality, conditional independence, constant variance, etc)
- Influence and leverage
- Collinearity and identifiability

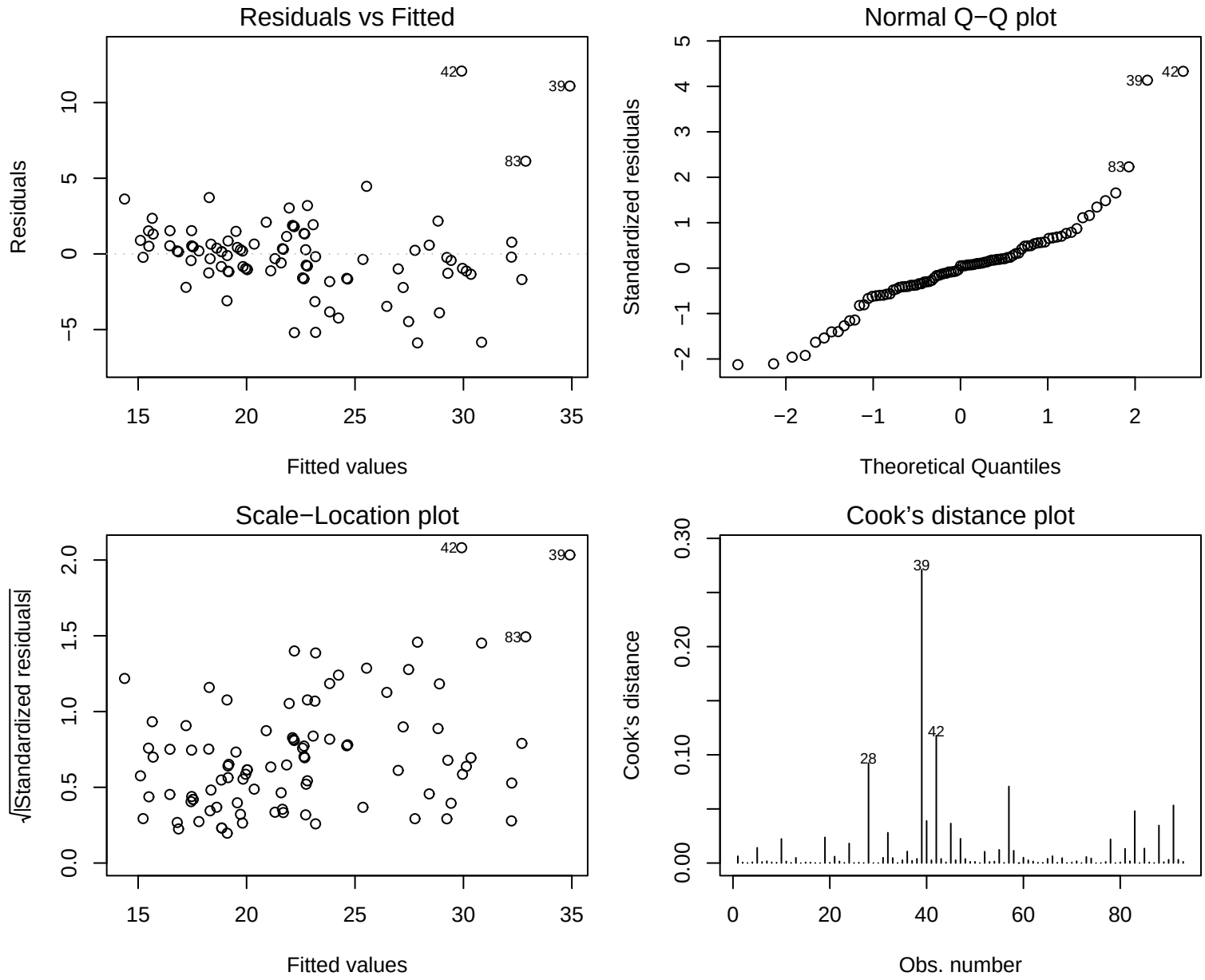
Based on the earlier plots and some theoretical considerations involving the physics of the situation, we will work with

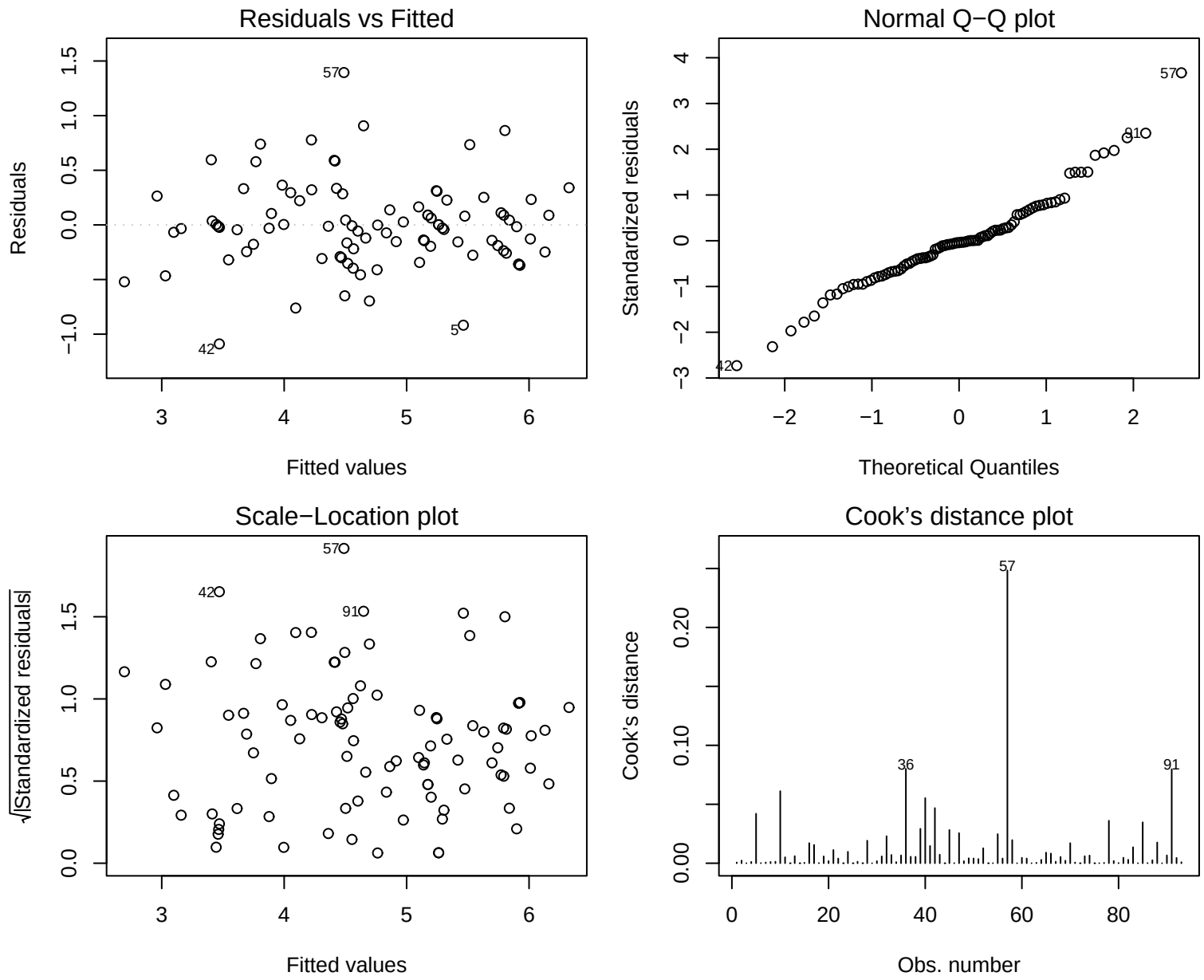
$$\text{CityFuel} = \frac{100}{\text{CityMPG}}$$

which is the number of gallons needed to go 100 miles as the response variable. This is also equivalent to how fuel use is reported in most countries.

This transformation helps with the linearity of the responses and the constant variance assumption of the deviations from the regression surface.


City Fuel
2
3
4
5
6
2000
2500
3000
3500
4000
Weight (lbs)
City Fuel
2
3
4
5
6
1
2
3
4
5
Engine Size (litres)
City Fuel
2
3
4
5
6
Compact
Midsize
Sporty
Van
Type
City Fuel
2
3
4
5
6
0
1
Domestic


Residuals vs Fitted
Residuals
Fitted values
42
39
83
Normal Q−Q plot
Standardized residuals
Theoretical Quantiles
Scale−Location plot
√|Standardized residuals|
Cook's distance plot
Cook's distance
Obs. number
28


Residuals vs Fitted
Residuals
Fitted values
57
42
5
Normal Q−Q plot
Standardized residuals
Theoretical Quantiles
57
91
42
Scale−Location plot
√|Standardized residuals|
Fitted values
57
42
91
Cook's distance plot
Cook's distance
Obs. number
57
36
91

The standard regression diagnostics look much better with the transformed response.

The flagged points for the two different regressions are

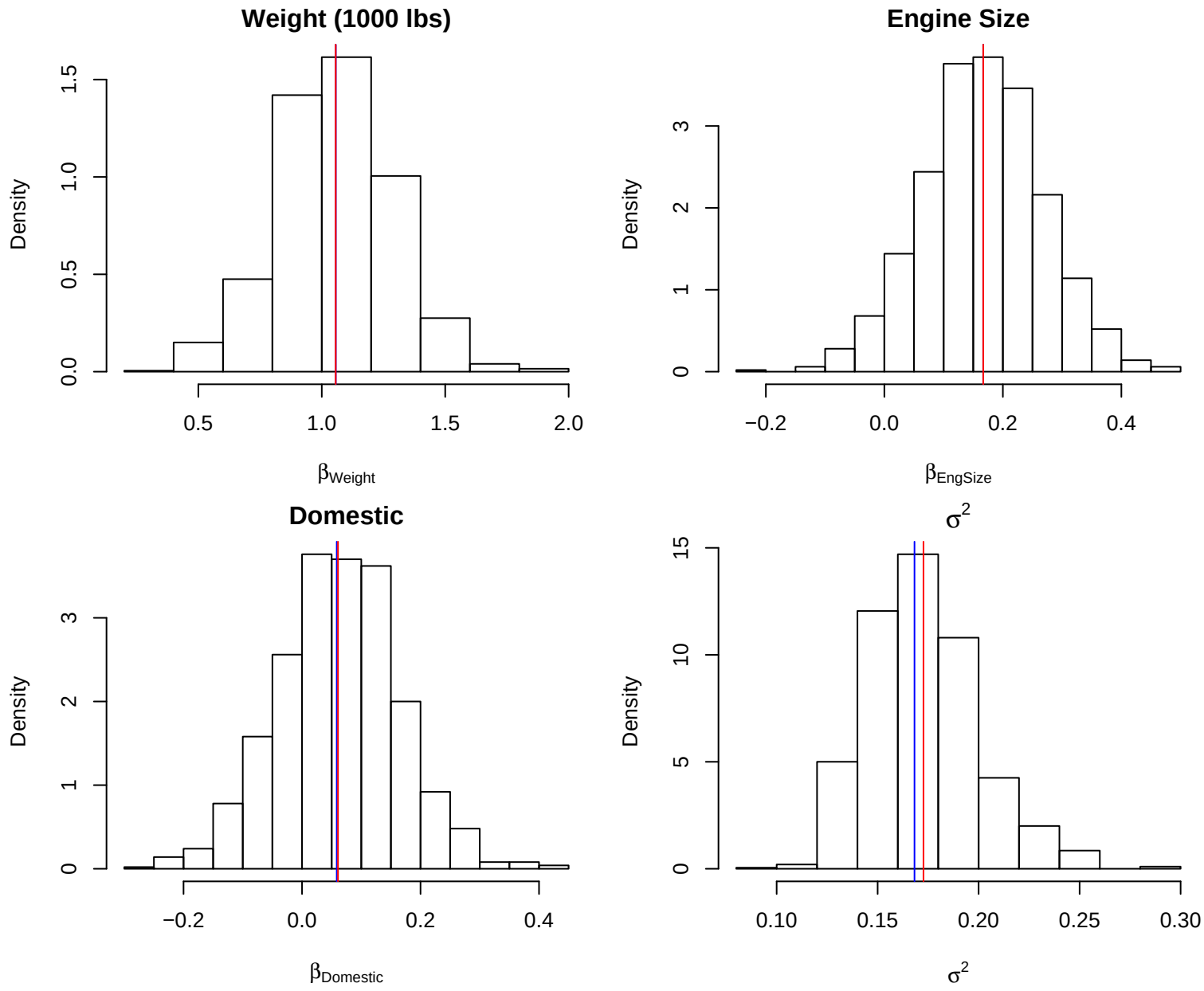
ID	Model	Weight	Engine Size	Type	Domestic	City MPG
5	BMW 535i	3640	3.5	Midsize	0	22
28	Dodge Stealth	3805	3.0	Sporty	1	18
36	Ford Aerostar	3735	3.0	Van	1	15
39	Geo Metro	1695	1.0	Small	0	46
42	Honda Civic	2350	1.5	Small	0	42
57	Mazda RX-7	2895	1.3	Sporty	0	17
83	Suzuki Swift	1965	1.3	Small	0	39
91	Volkswagen Corrado	2810	2.8	Sporty	0	18

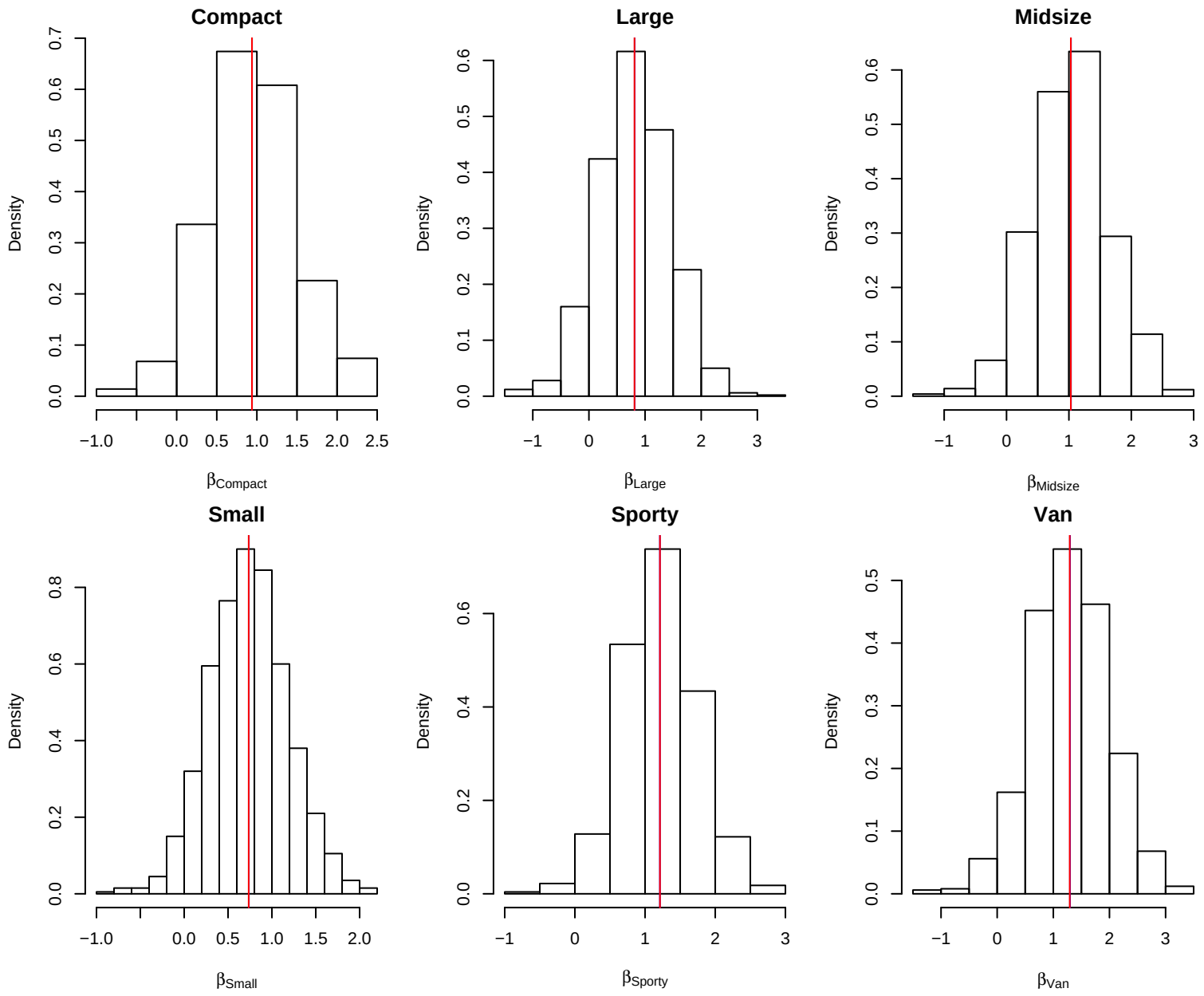
```
Call: lm(formula = CityFuel ~ Weight + EngSize + Type + Domestic - 1)

Residuals:
     Min       1Q   Median       3Q      Max
-1.08952 -0.23639 -0.01651  0.22142  1.39432

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
Weight       1.0568     0.2245      4.707  9.8e-06 ***
EngSize      0.1668307  0.0981565   1.700   0.0929 .
TypeCompact  0.9392962  0.5309963   1.769   0.0805 .
TypeLarge    0.8140547  0.6291158   1.294   0.1992
TypeMidsize  1.0327802  0.5953603   1.735   0.0865 .
TypeSmall    0.7367161  0.4352036   1.693   0.0942 .
TypeSporty   1.2116800  0.5243741   2.311   0.0233 *
TypeVan      1.2968046  0.6900831   1.879   0.0637 .
Domestic     0.0586004  0.0992826   0.590   0.5566

Residual standard error: 0.4102 on 84 degrees of freedom
Multiple R-Squared: 0.9934,     Adjusted R-squared: 0.9927
F-statistic:  1403 on 9 and 84 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: CityFuel
          Df  Sum Sq Mean Sq    F value  Pr(>F)
Weight     1 2122.41 2122.41 12615.1113 < 2e-16 ***
EngSize    1    0.04    0.04     0.2306 0.63234
Type       6    2.63    0.44     2.6004 0.02321 *
Domestic   1    0.06    0.06     0.3484 0.55662
Residuals 84   14.13    0.17
```

Notes:

- These are sequential SS thus you need to be careful in interpreting the F-tests.
- Weight has been adjusted to units of 1000 of pounds


Weight (1000 lbs)
Density
0.0
0.5
1.0
1.5
0.5
1.0
1.5
2.0
βWeight
Engine Size
Density
0
1
2
3
−0.2
0.0
0.2
0.4
βEngSize
Domestic
Density
0
1
2
3
−0.2
0.0
0.2
0.4
βDomestic
σ2
Density
0
5
10
15
0.10
0.15
0.20
0.25
0.30
σ2


Compact
Density
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
−1.0 0.0 0.5 1.0 1.5 2.0 2.5
βCompact
Large
Density
0.0 0.1 0.2 0.3 0.4 0.5 0.6
−1 0 1 2 3
βLarge
Midsize
Density
0.0 0.1 0.2 0.3 0.4 0.5 0.6
−1 0 1 2 3
βMidsize
Small
Density
0.0 0.2 0.4 0.6 0.8
−1.0 0.0 0.5 1.0 1.5 2.0
βSmall
Sporty
Density
0.0 0.2 0.4 0.6
−1 0 1 2 3
βSporty
Van
Density
0.0 0.1 0.2 0.3 0.4 0.5
−1 0 1 2 3
βVan

β	$E[\beta\|y]$	$\mathrm{SD}(\beta\|y)$	$P[\beta > 0\|y]$
β_{Weight}	1.056	0.233	1.000
β_{EngSize}	0.167	0.102	0.948
$\beta_{\mathrm{Domestic}}$	0.061	0.102	0.734
β_{Compact}	0.939	0.547	–
β_{Large}	0.812	0.640	–
β_{Midsize}	1.030	0.608	–
β_{Midsize}	0.735	0.447	–
β_{Sporty}	1.212	0.530	–
β_{Van}	1.293	0.699	–

$$E[\sigma^2|y] = 0.173 \qquad \mathrm{SD}(\sigma^2|y) = 0.027$$

$$E[\sigma|y] = 0.414 \qquad \mathrm{SD}(\sigma|y) = 0.032$$

$P[\beta_{\text{Row}} > \beta_{\text{Column}}|y]$

	β_{Compact}	β_{Large}	β_{Midsize}	β_{Small}	β_{Sporty}	β_{Van}
β_{Compact}	–	0.733	0.283	0.890	0.032	0.060
β_{Large}	0.267	–	0.102	0.623	0.029	0.015
β_{Midsize}	0.717	0.898	–	0.930	0.134	0.079
β_{Small}	0.110	0.377	0.070	–	0.003	0.031
β_{Sporty}	0.968	0.971	0.866	0.997	–	0.366
β_{Van}	0.940	0.985	0.921	0.969	0.634	–