pdf_draft

Kaleb Cervantes

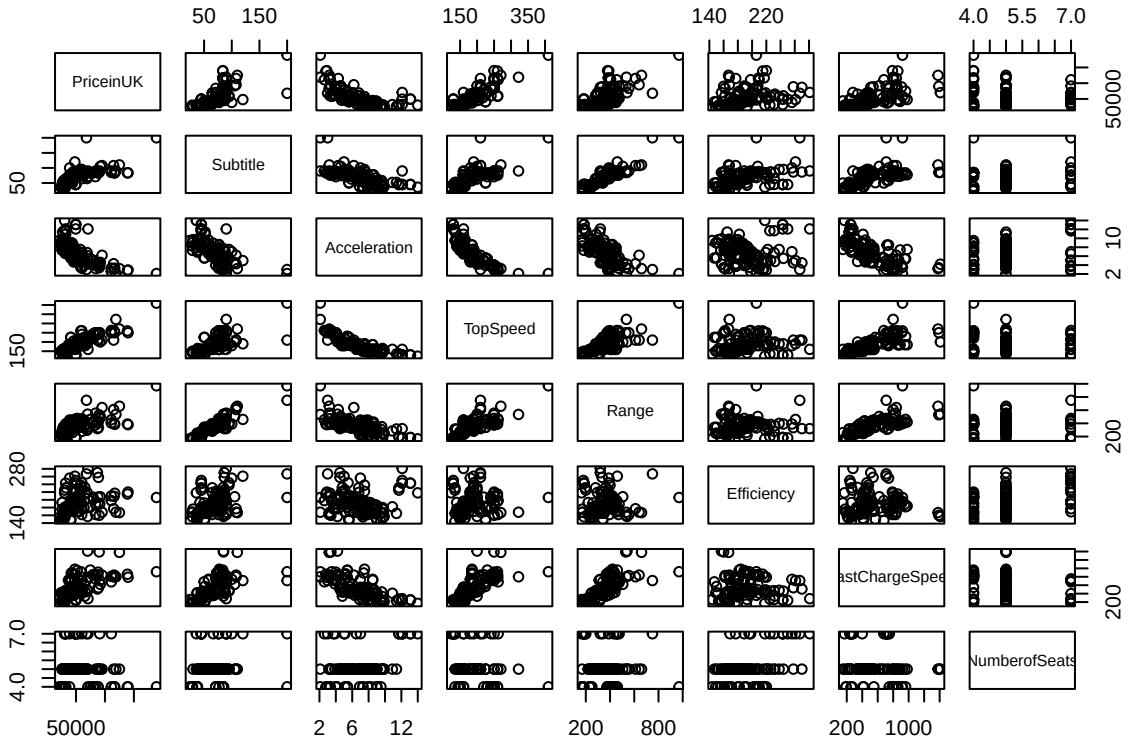
4/20/2022

Model Selection

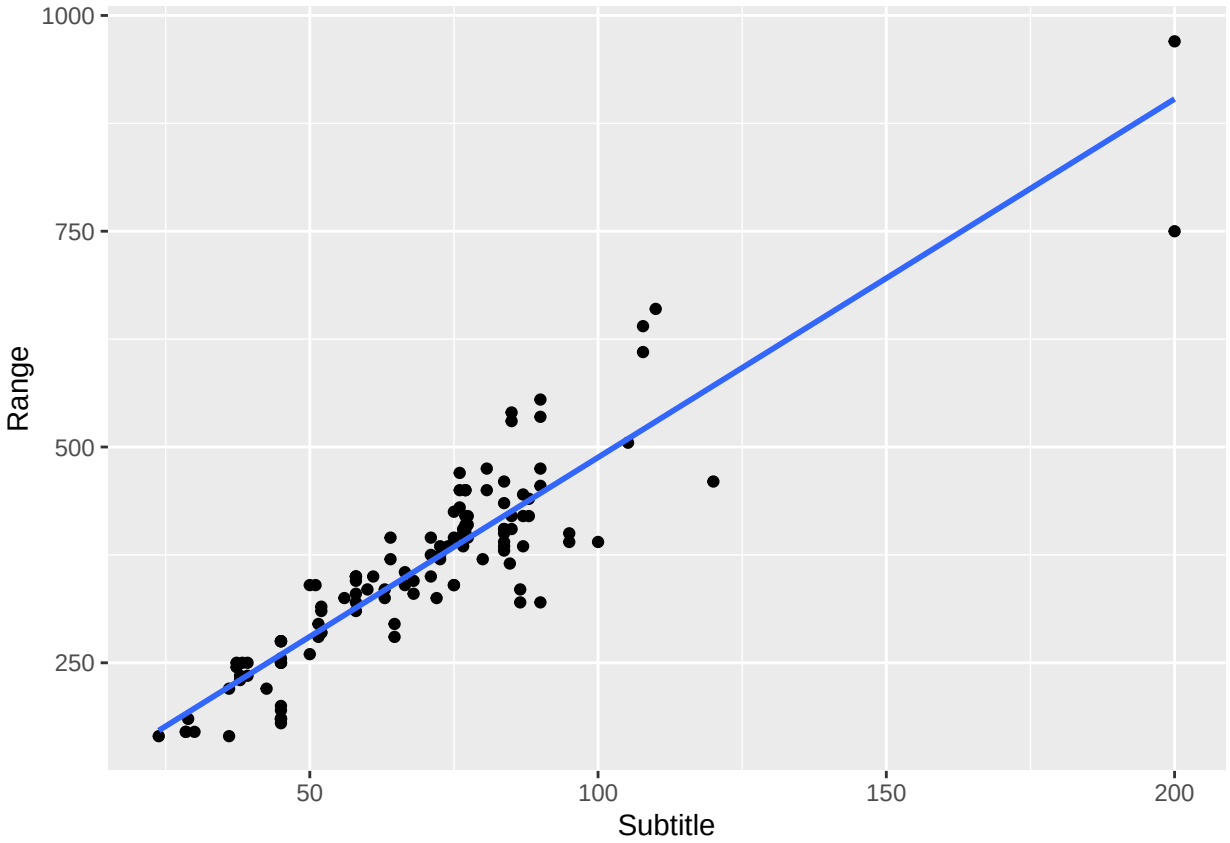
Variable Selection

The first thing was to check the correlations between some of the variables and visualize the data.

```
##                    Subtitle Acceleration    TopSpeed       Range  Efficiency
## Subtitle         1.00000000  -0.65205890  0.68551542  0.91104205  0.42266885
## Acceleration    -0.65205890   1.00000000 -0.83536291 -0.70220767 -0.03009386
## TopSpeed         0.68551542  -0.83536291  1.00000000  0.74349641  0.07810845
## Range            0.91104205  -0.70220767  0.74349641  1.00000000  0.04496225
## Efficiency       0.42266885  -0.03009386  0.07810845  0.04496225  1.00000000
## FastChargeSpeed  0.60230194  -0.71134894  0.73572843  0.72271112 -0.02642250
## NumberofSeats    0.08103089   0.30861901 -0.21958021 -0.06531154  0.44090323
## PriceinUK        0.70390802  -0.70863967  0.86502374  0.68439993  0.30531728
##                  FastChargeSpeed NumberofSeats  PriceinUK
## Subtitle               0.6023019    0.08103089  0.7039080
## Acceleration          -0.7113489    0.30861901 -0.7086397
## TopSpeed               0.7357284   -0.21958021  0.8650237
## Range                  0.7227111   -0.06531154  0.6843999
## Efficiency            -0.0264225    0.44090323  0.3053173
## FastChargeSpeed        1.0000000   -0.16516351  0.6495690
## NumberofSeats         -0.1651635    1.00000000 -0.1491998
## PriceinUK              0.6495690   -0.14919979  1.0000000
```



The most highly correlated variables are `Subtitle` and `Range`. These have a correlation of 0.91104205 and seem to have similar plots in the scatter plot matrix. A zoomed in scatter plot is included below.



When zoomed in, the two variables seem to have a linear relationship with each other. This indicates that one of these may be dropped. `Subtitle` and `Range` have respective correlations 0.7039080 and 0.6843999 with the response. Since `Subtitle` has a stronger correlation with the response, that will be the predictor that is kept.

Transformations

In the first row of the scatter plot matrix, `TopSpeed` and `Acceleration` appear to form a parabola when next to each other. As such, a quadratic transformation will be applied to those two variables.

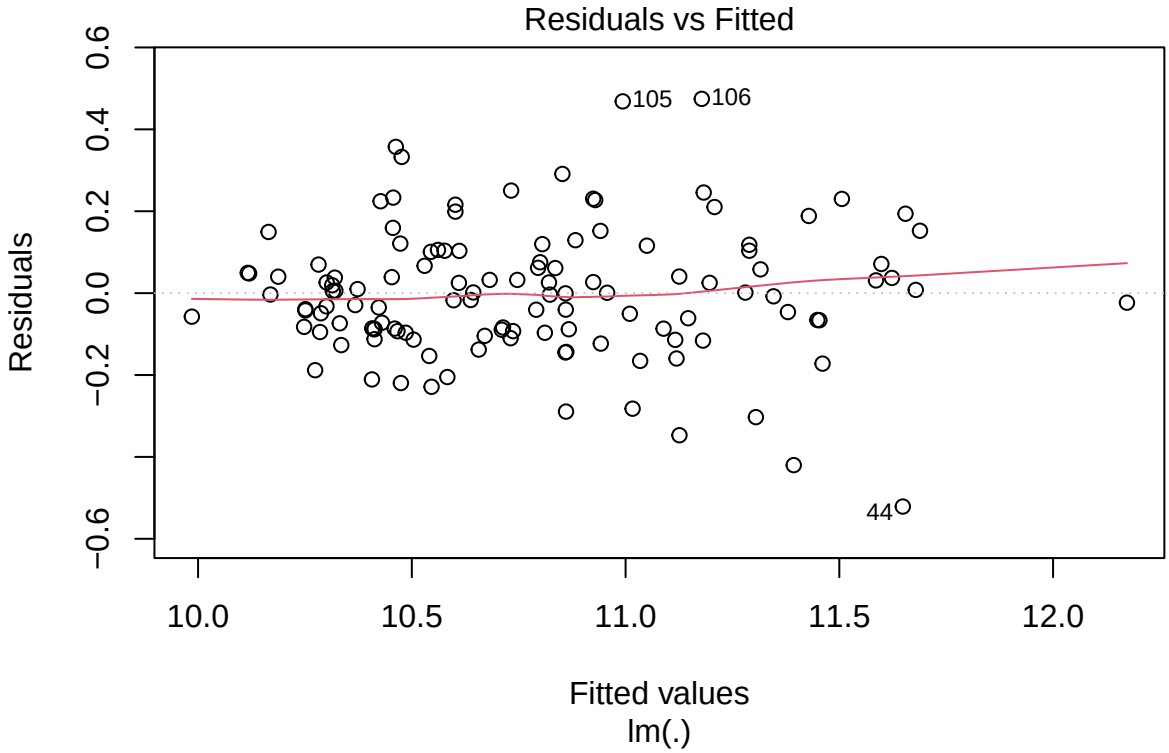
`NumberofSeats` also appears to behave as a factor in the scatter plot matrix. As such, it will be transformed into one.

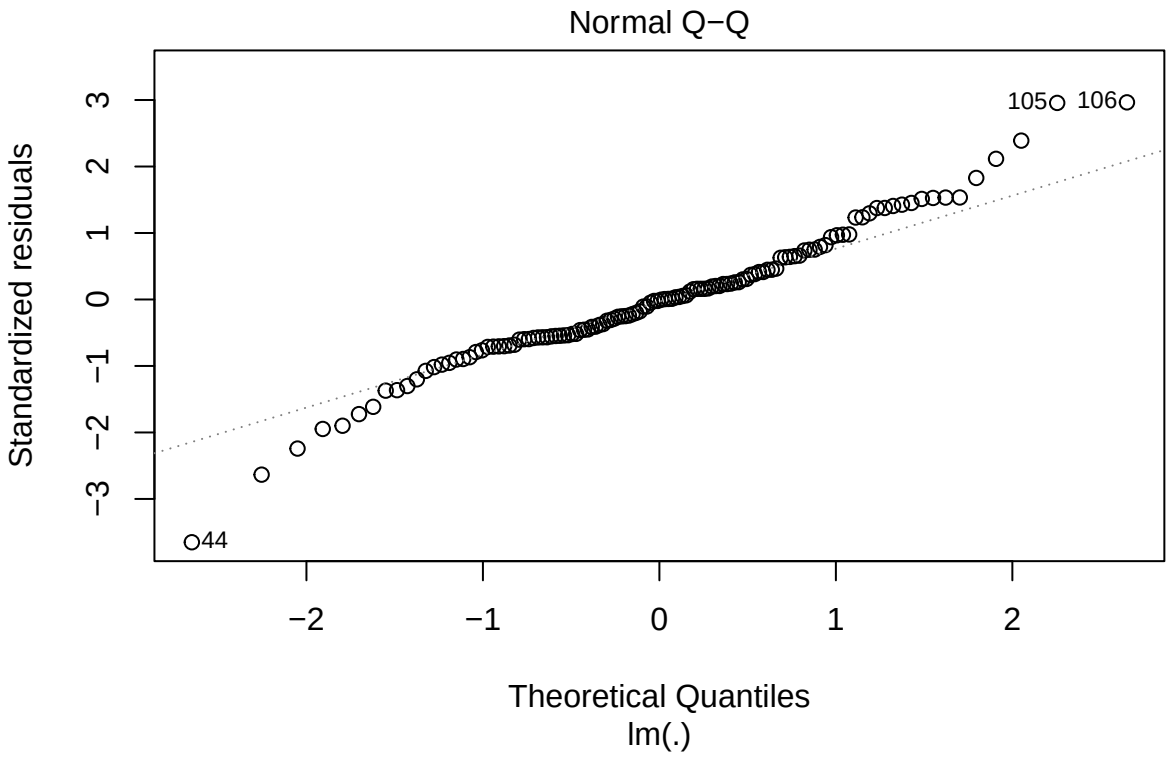
In order to handle non-constant variance, logarithmic transformations will be applied to the other predictors and response variables.

Full Model

```
##
## Call:
## lm(formula = ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.52123 -0.09026 -0.00196  0.08179  0.47444
##
## Coefficients:
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.237e+00  9.592e-01   5.460 2.90e-07 ***
## log(Subtitle)                   4.012e-01  9.933e-02   4.039 9.86e-05 ***
## poly(Acceleration, 2, raw = T)1 -2.076e-01  5.572e-02  -3.726 0.000307 ***
## poly(Acceleration, 2, raw = T)2  1.387e-02  3.076e-03   4.510 1.61e-05 ***
## poly(TopSpeed, 2, raw = T)1      1.321e-02  3.176e-03   4.159 6.30e-05 ***
## poly(TopSpeed, 2, raw = T)2     -1.910e-05  5.803e-06  -3.291 0.001334 **
## log(Efficiency)                  5.679e-01  1.738e-01   3.267 0.001443 **
## log(FastChargeSpeed)            -6.657e-03  6.781e-02  -0.098 0.921973
## DriveFront Wheel Drive          -2.338e-02  6.980e-02  -0.335 0.738269
## DriveRear Wheel Drive           -3.726e-02  5.891e-02  -0.632 0.528352
## factor(NumberofSeats)5          -1.268e-01  4.791e-02  -2.646 0.009308 **
## factor(NumberofSeats)7          -2.770e-01  7.294e-02  -3.798 0.000238 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1665 on 112 degrees of freedom
## Multiple R-squared:  0.8815, Adjusted R-squared:  0.8698
## F-statistic: 75.71 on 11 and 112 DF,  p-value: < 2.2e-16
```



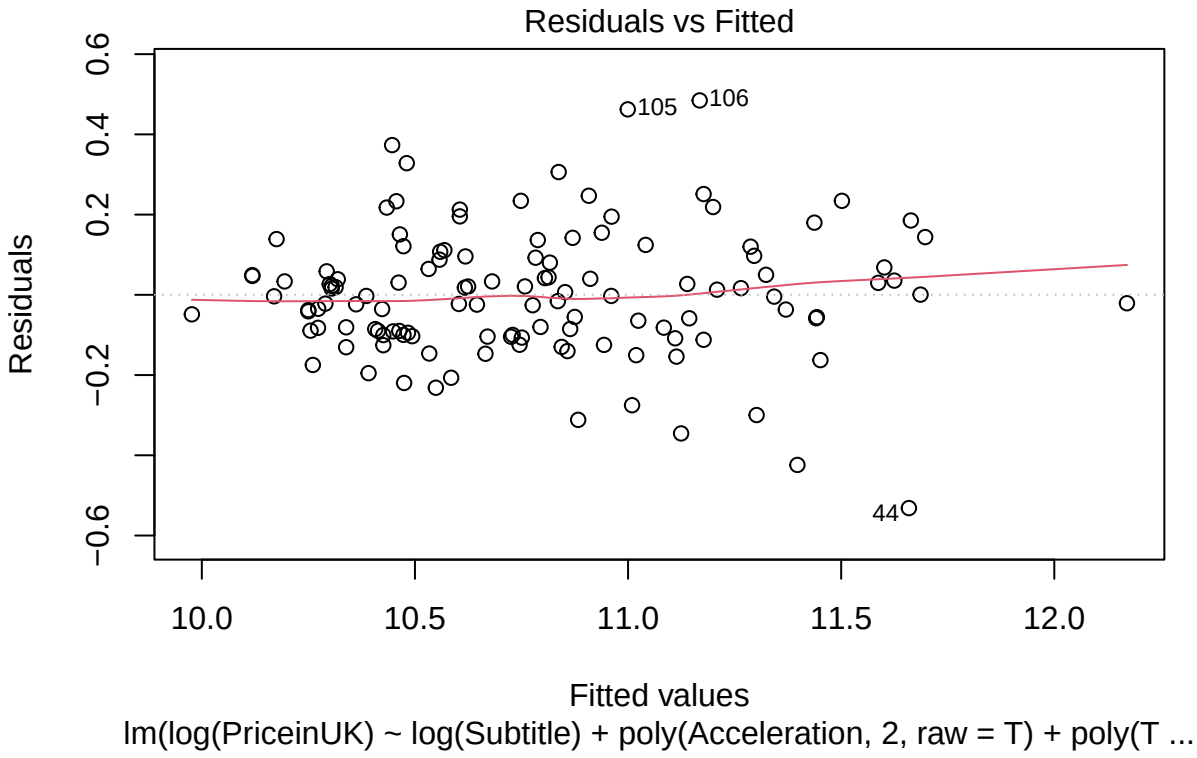


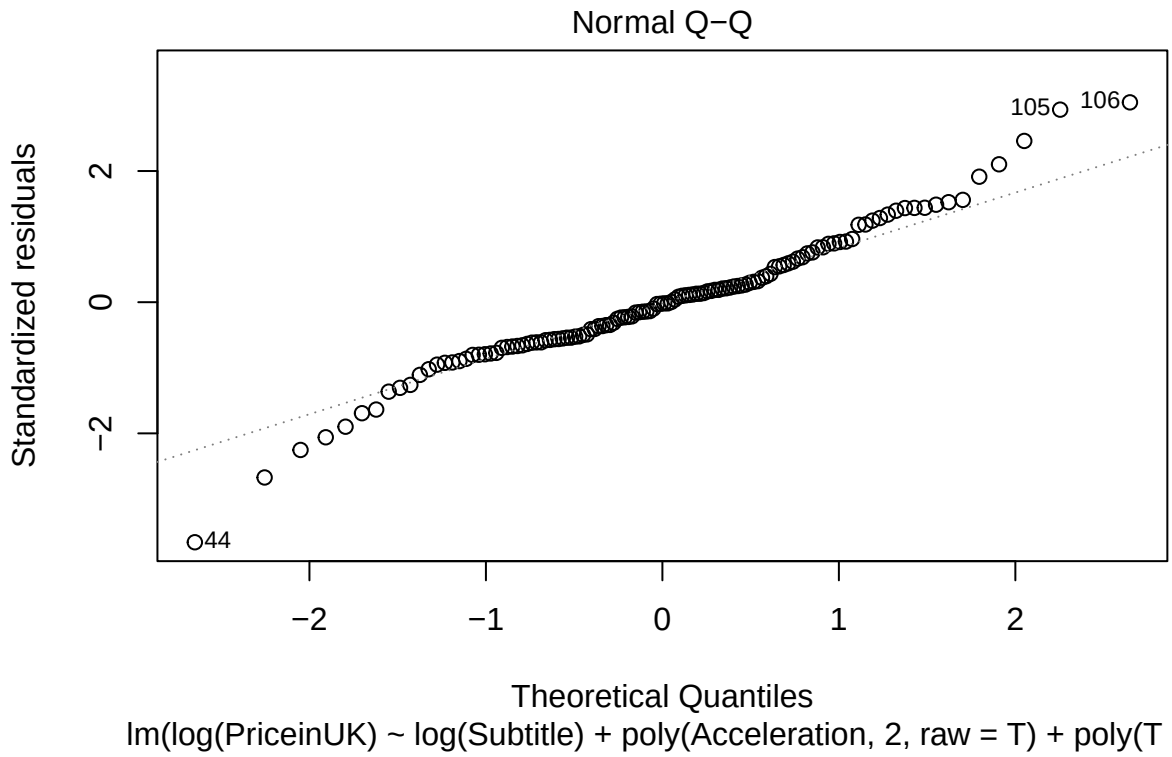
Reduced Model

This full model seems to have met the assumptions for linear regression. However, it still contains predictors that are not significant. In order to reduce the model, a step-wise process will be used with BIC as the metric.

```
##
## Call:
## lm(formula = log(PriceinUK) ~ log(Subtitle) + poly(Acceleration,
##     2, raw = T) + poly(TopSpeed, 2, raw = T) + log(Efficiency) +
##     factor(NumberofSeats), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.53164 -0.09592 -0.00326  0.08894  0.48461
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.139e+00  8.158e-01   6.299 5.67e-09 ***
## log(Subtitle)                    3.836e-01  8.049e-02   4.766 5.55e-06 ***
## poly(Acceleration, 2, raw = T)1 -2.248e-01  4.879e-02  -4.608 1.06e-05 ***
## poly(Acceleration, 2, raw = T)2  1.456e-02  2.845e-03   5.117 1.26e-06 ***
## poly(TopSpeed, 2, raw = T)1      1.280e-02  2.711e-03   4.722 6.66e-06 ***
## poly(TopSpeed, 2, raw = T)2     -1.852e-05  5.013e-06  -3.695 0.000338 ***
## log(Efficiency)                  6.144e-01  1.454e-01   4.225 4.81e-05 ***
## factor(NumberofSeats)5          -1.197e-01  4.577e-02  -2.615 0.010118 *
## factor(NumberofSeats)7          -2.726e-01  7.175e-02  -3.799 0.000234 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1646 on 115 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8727
## F-statistic: 106.4 on 8 and 115 DF,  p-value: < 2.2e-16
```





This process removed `log(FastChargeSpeed)` and `Drive` from the model. The assumptions still appear to be met. In order to ensure no significant predictors were removed, a partial F-Test will be conducted between the two models.

```
## Analysis of Variance Table
##
## Model 1: log(PriceinUK) ~ log(Subtitle) + poly(Acceleration, 2, raw = T) +
##     poly(TopSpeed, 2, raw = T) + log(Efficiency) + factor(NumberofSeats)
## Model 2: log(PriceinUK) ~ log(Subtitle) + poly(Acceleration, 2, raw = T) +
##     poly(TopSpeed, 2, raw = T) + log(Efficiency) + log(FastChargeSpeed) +
##     Drive + factor(NumberofSeats)
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1    115 3.1166
## 2    112 3.1036  3  0.012972 0.156 0.9256
```

Since the p-value is high, we lack statistically significant evidence that any significant predictors were dropped.

This leaves us with the formula:

$$
\begin{aligned}
\ln \texttt{PriceinUK} = \beta_0 & \\
& + \beta_1 \ln \texttt{Subtitle} \\
& + \beta_2 \texttt{Acceleration} + \beta_3 \texttt{Acceleration}^2 \\
& + \beta_4 \texttt{TopSpeed} + \beta_5 \texttt{TopSpeed}^2 \\
& + \beta_6 \ln \texttt{Efficiency} \\
& + \beta_7 \left(\texttt{NumberofSeats} == 5\right) + \beta_8 \left(\texttt{NumberofSeats} == 7\right) \\
& + \epsilon
\end{aligned}
$$