Regression Models Course Project

Paul Clark: March 5, 2017

1 Executive Summary

For its 1974 edition, US magazine *Motor Trend* has asked two questions to be addressed using data on fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-74 models). **Question 1:** *"Is an automatic or manual transmission better for MPG?"* **Question 2:** *"How does MPG differ, quantitatively, between automatic and manual transmissions?"* The variables are described below:

```
 [1] "mpg  Miles/(US) gallon"          "cyl  Num cylinders"
 [3] "disp Displacement (cu.in.)"      "hp   Gross horsepower"
 [5] "drat Rear axle ratio"            "wt   Weight (1000 lbs)"
 [7] "qsec 1/4 mile time"              "vs   Engine(0=V,1=Inline)"
 [9] "am   Transmission(0=auto,1=man)" "gear Num forward gears"
[11] "carb Num carburetors"
```

We address the questions via two approaches: **Approach (a)** - *Calculation of mean differenc and two-sample T-test on `mpg` of the transmission types.* **Approach (b)** - *Adjustment of the `am` effect in (a) by regressing `mpg` on all other variables in the data.* Note: for either approach to be meaningful, this small sample of 32 cars must be representative of their populations. We assume this is the case.

We conclude from **Approach (a)** that the two groups of cars come from populations with statistically different means. If the data provided is representative, then *manual* transmission cars generally have an estimated mean gas mileage **7.2 MPG** higher than *automatic* cars.

From **Approach (b)**, we have two main conclusions: **(1)** the adjustments of the `am` effect that most help predict `mpg` are `hp` and `wt`, both negative, corresponding to model $\mathbf{E[mpg]} = \mathbf{M} \cdot am + \mathbf{H} \cdot hp + \mathbf{W} \cdot wt$; **(2)** the *automatic/manual* effect M, with *horsepower* and *weight* held fixed, is approximately **2.1 MPG**, with manual again having higher `mpg` than automatic. Other engineered effects are the sources of more of the difference in `mpg` (**5 MPG** worth) than transmission type alone. Increased engine `hp` accounts for a decrease of 3.7 MPG per hundred HP, and increased weight accounts for a decrease of 2.9 MPG per thousand lbs.

Finally, due to low significance of the `am` coefficient (**Std. Error** of 1.4), we also investigated other models not constrained to include `am`. The best of these was $\mathbf{E[mpg]} = \mathbf{C} \cdot cyl + \mathbf{W} \cdot wt$, with effect sizes of $C =$ 1.5 MPG decrease per cylinder, and $W = 3.2$ MPG decrease per thousand lbs of weight. Note the rough consistency of the `wt` effect across the two models: it is again around 3 MPG.

2 Exploratory Data Analysis

In **Figure 1**, we examine integer predictors to decide whether to treat them as factors. We find value in treating `cyl` and `carb` as continuous: they show clear trends vs. other variables. And from a pairs plot, **Figure 2**, we see many strong correlations, so model selection should consider variance inflation.

3 Approach (a): Two sample t-test and Inference

Figure 3 shows that in this data, `mpg` varies with `am`. Mean mileage for manual is **7.2 mpg** higher than automatic. We compute p-value and confidence interval for the test of manual transmission *greater*.

```
t_test<-t.test(mpg~am,data=mtcars,alternative='less') # less: 2nd factor is t.test base
# note: code for processing and formatting of output suppressed
```

```
p-value = 0.07%     95% conf.int = 3.91 to Inf
```

Inference: Given the p-value, we are highly confident manual is associated with higher fuel economy, in the populations from which these samples were drawn.

4 Approach (b): OLS regression

Approach **(b)** is under-specified. Having identified `mpg` and `am` as of interest, the "correct" choice of model still depends on selection of the appropriate subset of 9 other variables. This should be a function of variable/model significance, but also *Motor Trend's* interests. For significance testing, manual checking of p-values is in-viable: at least $2^9 = 512$ models with single predictors exist. Therefore, to fully specify and make the approach manageable, we must make additional assumptions.

We assume *Motor Trend* values: **(A)** Parsimony/simplicity; **(B)** Models with granular, causal variables that may clarify engineering trade-offs; **(C)** Predictiveness: good generalizability outside the training sample.

4.1 Model Search

Due to **(A)**, we consider no interactions. From **(B)**, we exclude `qsec`, a summary metric. Due to **(C)**, we rank models using the **AIC** metric, which estimates model predictiveness outside the training sample. For OLS regression, the metric is $n \cdot Log(\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}) + 2k$, where $k = \#$ of parameters including estimate of residual error, an overfitting penalty. In fact, we use **AICc**, which corrects the penalty to be greater for small n by adding a term $\frac{2k(k+1)}{n-k-1}$ (note that this term varies based on model structure; this version only holds for Gaussian models). We use automated search to make evaluation of all 2^9 models feasible. Models are ranked from smallest to largest AICc. Only non-zero coefficients are shown, and only models with the variable of interest (`am`) are evaluated.

```
if (!"MuMIn" %in% row.names(installed.packages())) {install.packages("MuMIn")}
library(MuMIn); mtcars$qsec <- NULL; mtcars$gear <- as.factor(mtcars$gear)
globalmodel <- lm(mpg ~ ., data = mtcars, na.action = na.fail)
bestmodels <- dredge(globalmodel, subset = ~ am) # only considers models with `am`
bestmodels[1:5,]
```

\	(Int)	am	carb	cyl	hp	vs	wt	df	logLik	AICc	delta	weight
322	34.0	2.08			-0.037		-2.9	5	-73.1	158.4	0.0	0.31
264	36.9	1.78	-0.75	-1.20			-2.5	6	-72.0	159.4	0.9	0.20
450	31.1	2.42			-0.030	1.8	-2.6	6	-72.0	159.4	1.0	0.19
326	36.1	1.48		-0.75	-0.025		-2.6	6	-72.1	159.6	1.2	0.17
262	39.4	0.18		-1.51			-3.1	5	-74.0	160.3	1.9	0.12

4.2 Model Inference & Interpretation of Coefficients

We investigate values of the `am` coefficient for the top models. Note the first 3 all round to 2, suggesting this is a good rough estimate of the adjusted transmission effect. Although our top model is only one AICc point lower than the next best model (model averaging is suggested via the weights, for differences less than 2), we focus attention on it, in the spirit of Assumption **(A)**.

```
             Estimate Std. Error
(Intercept)   34.003      2.643
am             2.084      1.376
hp            -0.037      0.010
wt            -2.879      0.905
```

The R^2 of this top model is 84%: it explains a high degree of sample variance with only 3 covariates. The above table contains no p-values, as after a search of 2^9 models, these would be inflated: since the procedure only selects 'good' models for consideration, we need to control the "False Discovery Rate", which is the fraction of **all rejected** null hypotheses which are false (i.e., $(False\ Positives)/(False\ Positives + True\ Positives)$), not just α, which is the fraction of **all truly 0** results that are rejected (i.e., $(False\ Positives)/(False\ Positives + True\ Negatives)$). However, standard errors are provided. These show `hp` (with 4 MPG decrease per 100 HP increase) and `wt` (with 3 MPG decrease per 1000 lb increase) are significant, whereas significance of the the `am` coefficient, 2.1, is low. The *Estimate* over the *Std. Error*, or t-stat, is only 1.5. Two is near the $\alpha = 5\%$ threshold. Given this, we investigate models that do **not** include `am`. Here are the top 10 overall:

```r
best_unconmodels <- dredge(globalmodel) # considers 'uncon'-strained models
best_unconmodels[1:10,]
```

\	(Int)	am	carb	cyl	disp	drat	hp	vs	wt	df	logLik	AICc	dlta	wgt
261	39.7			-1.5					-3.2	4	-74.0	157.5	0.0	.17
325	38.8			-0.9			-.018		-3.2	5	-72.7	157.8	0.3	.15
263	39.6		-.49	-1.3					-3.2	5	-72.8	157.9	0.4	.14
321	37.2						-.032		-3.9	4	-74.3	158.1	0.6	.12
322	34.0	2.1					-.037		-2.9	5	-73.1	158.4	1.0	.10
337	29.4					1.6	-.032		-3.2	5	-73.4	159.0	1.5	.08
264	36.9	1.8	-.75	-1.2					-2.5	6	-72.0	159.4	1.9	.07
450	31.1	2.4					-.030	1.8	-2.6	6	-72.0	159.4	1.9	.06
326	36.1	1.5		-0.7			-.025		-2.6	6	-72.1	159.6	2.1	.06
333	40.8			-1.3	.012		-.021		-3.9	6	-72.2	159.7	2.2	.06

We see above that models involving `am` do not appear until rank 5 and below. However, 3 of these occur at *delta* < 2, generally considered to be within the margin of error. But, given **(A)**, it is prudent to add the top model overall, involving only `cyl` and `wt`, to the results presented to *Motor Trend*. Compared to the model containing `am`, it has one fewer parameter - therefore less likely to fall victim to overfitting, and only slightly lower R^2, equal to 83%. Also, the ratios of *Std. Errors* to *Estimates* make all coefficients appear significant. This model implies, though, that none of the variance in MPG is really due to transmission type, but to the combined effect of # of cylinders and weight.

```
            Estimate Std. Error
(Intercept)    39.69       1.71
cyl            -1.51       0.41
wt             -3.19       0.76
```

4.3 Model Diagnostics

We run base R's standard plots in **Figure 4**. Though the smoother line in *Residuals vs Fitted* shows curvature, pointing to possible quadratic terms, the trend is not pronounced except for the 3 labeled points (*Toyota Corolla, Fiat 128,* and *Chrysler Imperial*). These have notably higher MPG than the trend. *Normal Q-Q* shows a somewhat right skewed distribution beyond 1 normal quantile. But the deviations are not extreme, except for the 3 labeled points, and the lowest. The lowest, given by the code below, is *Mazda RX4*.

```r
bestmodel<-lm(mpg~am+hp+wt,mtcars); row.names(mtcars)[which.min(bestmodel$residuals)]
```

Scale-Location shows some heteroskedasticity. In *Residuals vs Leverage*, all points are inside 0.5 *Cook's distance*, indicating stable βs. Finally, from package `car`, we use `vif()` to calculate the Variance Inflation Factors and evaluate collinearity. All are under 5, causing no alarm (code suppressed to conserve space). VIFs:

```
  am   hp   wt
2.27 2.09 3.77
```

Given, especially, the heteroskedasticity, we examine the diagnostics for the top model overall, too (**Figure 5**, `mpg ~ cyl + wt`). It does not appear markedly better anywhere, and it appears slightly worse on the **Normal Q-Q** evaluation.

5 Figures

Plots provide guidance on whether to treat integer variables as continuous or factor.

```r
data(mtcars)
intplotsdf <- data.frame(mpg=mtcars$mpg, hp=mtcars$hp, cyl=mtcars$cyl, carb=mtcars$carb,
                 gear=mtcars$gear, qsec=mtcars$qsec)
if (!"tidyr" %in% rownames(installed.packages())) {install.packages("tidyr")}
library(tidyr)
intplotsdf <- gather(data = intplotsdf, key = x_variable,
                       value = x_value, -mpg, -hp, -qsec)
intplotsdf <- gather(data = intplotsdf, key = y_variable,
```

```
                            value = y_value, -x_variable, -x_value)
if (!"ggplot2" %in% rownames(installed.packages())) {install.packages("ggplot2")}
library(ggplot2)
g_integer_vars <- ggplot(intplotsdf, aes(x=x_value,y=y_value)) +
                        facet_grid(y_variable ~ x_variable, scales = "free") +
                   geom_point() + geom_smooth(method = "lm") +
labs(title = "'gear' behaves as a factor, 'carb' and 'cyl' can be treated as continous")
print(g_integer_vars)
```

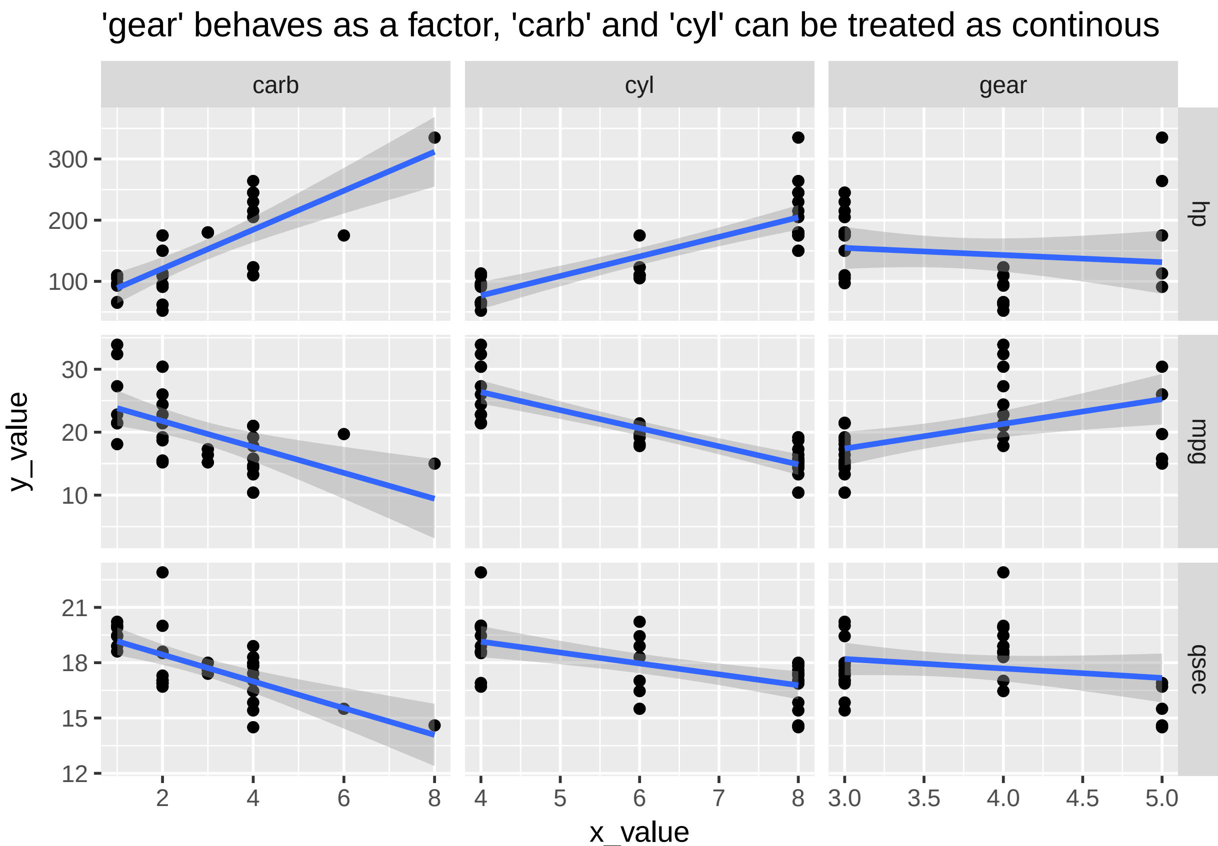

Figure 1: Plot of continuous vs. factor variables in mtcars data

The pairs plot shows many strong correlations.

```
if (!"GGally" %in% rownames(installed.packages())) {install.packages("GGally")}
library(GGally)
g_pairs <- ggpairs(mtcars, mapping=aes(color=am, alpha = 0.7),lower = list(continuous =
                                                            wrap(ggally_smooth, size = 1)),
            diag = list(continuous = "barDiag"), upper = list(continuous =
            wrap(ggally_cor,size=2)), axisLabels = 'none')

print(g_pairs)
```

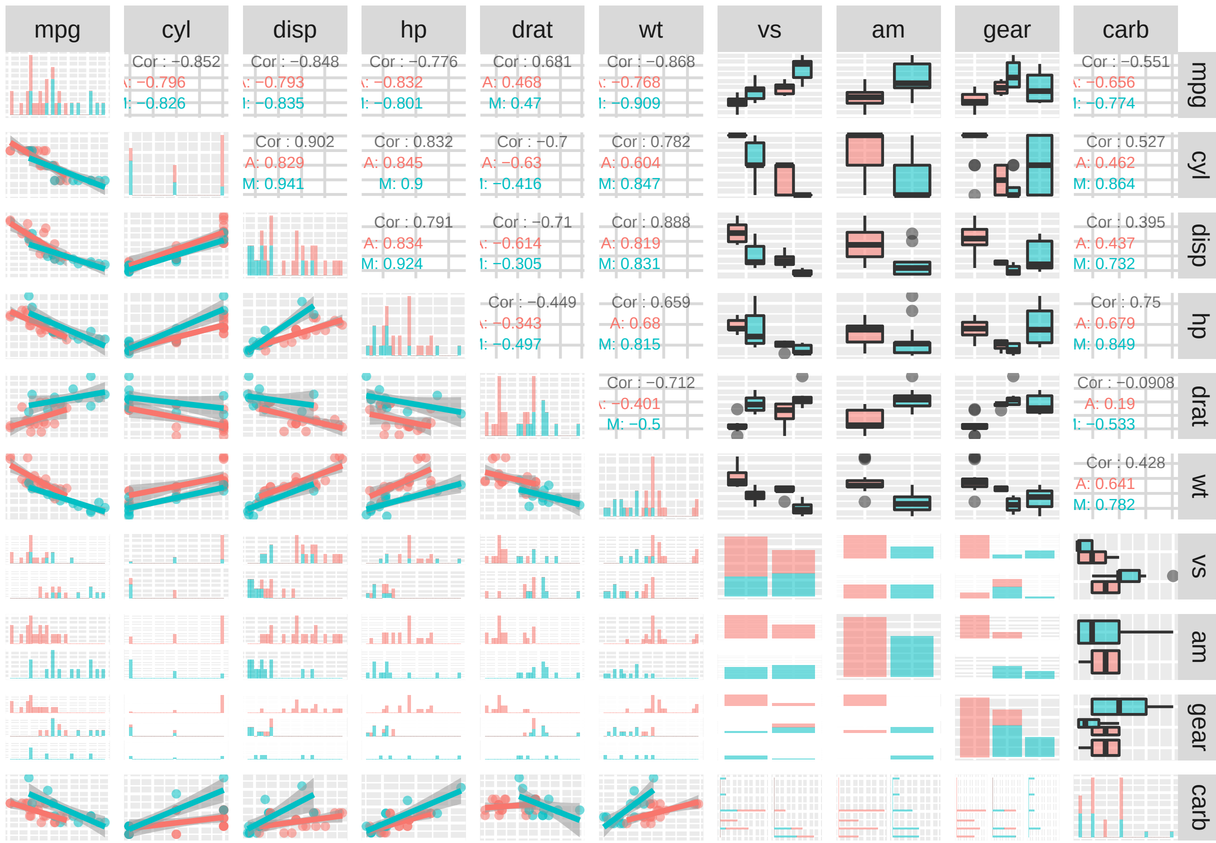

Figure 2: Pairs Plot of Motor Trend Cars Database

The violin plot depicts strong association between transmission type and `mpg`.

```
g_violin <- ggplot(mtcars, aes(x = am, y = mpg)) + geom_violin() +
        geom_dotplot(binaxis='y', stackdir='center', dotsize=1) +
        stat_summary(fun.y=mean, geom="point", shape=23, size=4, aes(fill = am)) +
        scale_fill_discrete(name="Mean MPG")+
        labs(title = "Average MPG for manual transmissions is significantly higher")
print(g_violin)
```

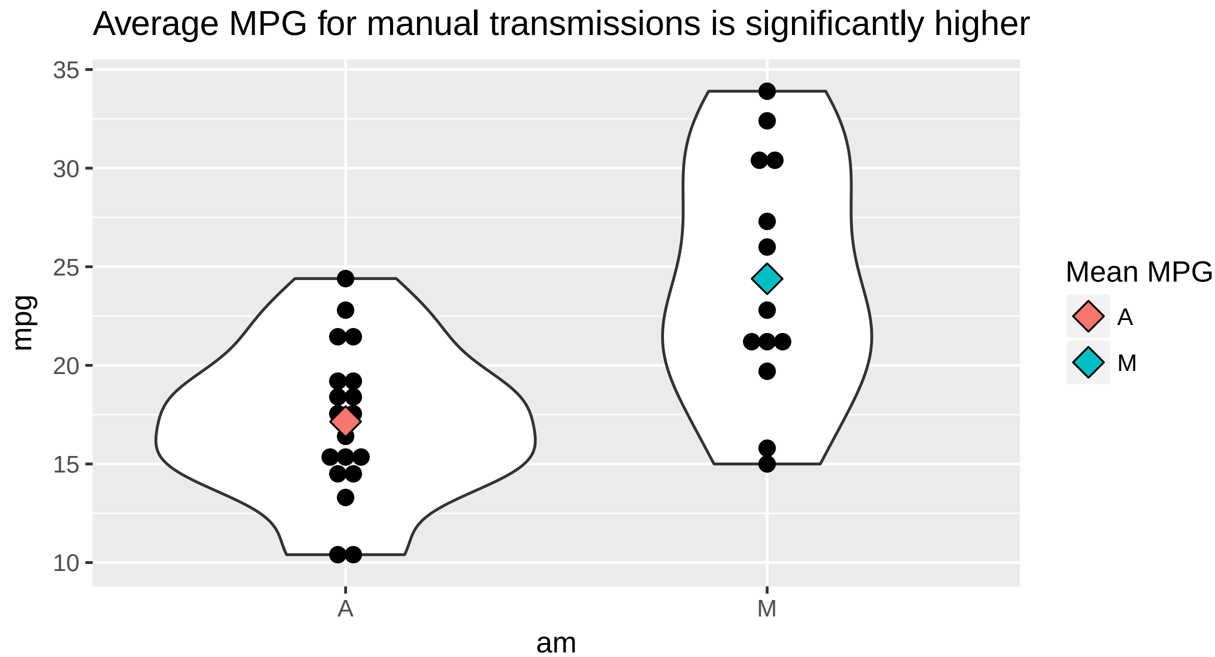

Figure 3: Violin plot of MPG vs. Transmission type

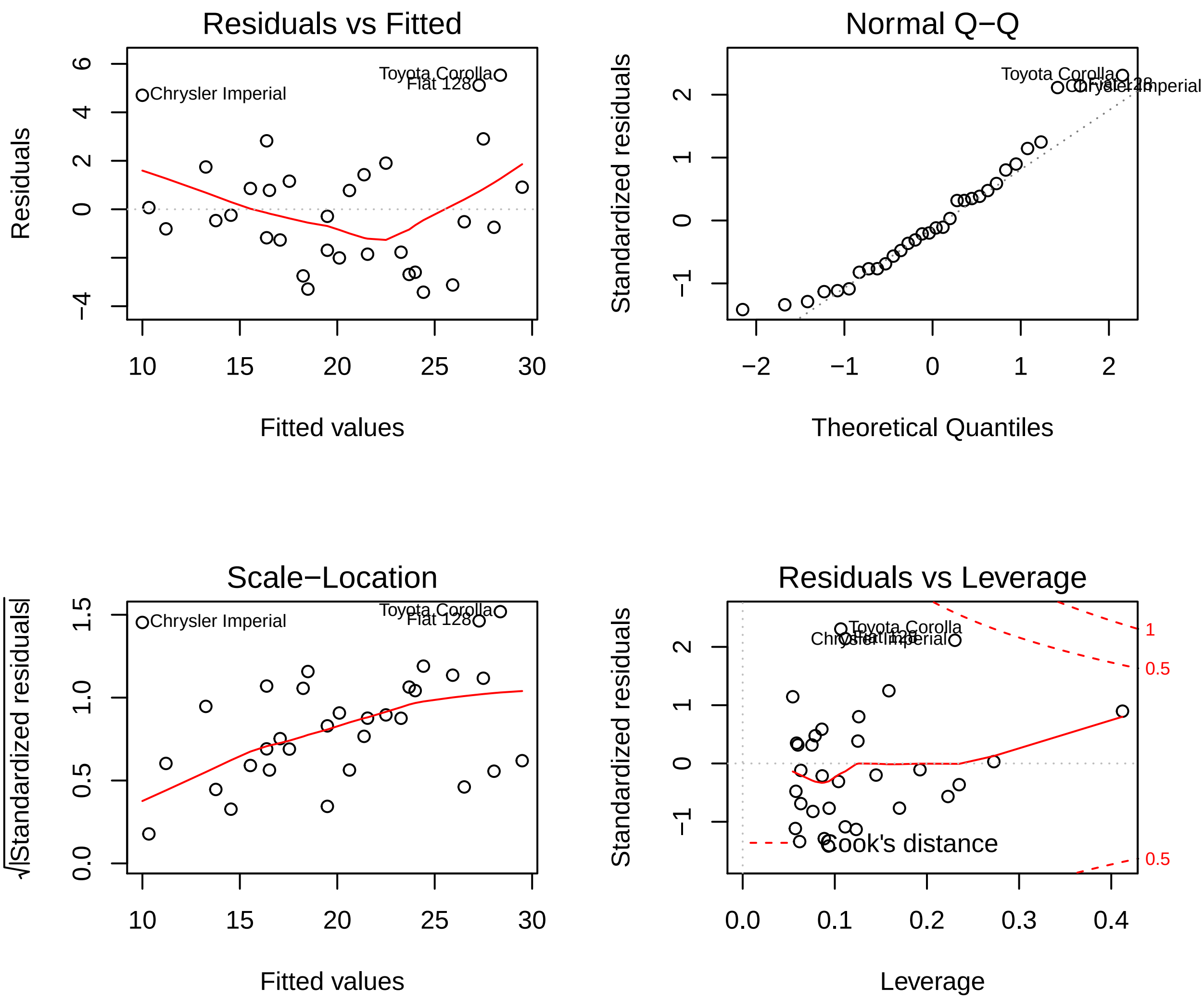

Figure 4: Base R’s standard diagnostic plots guide our eye to potential problems.

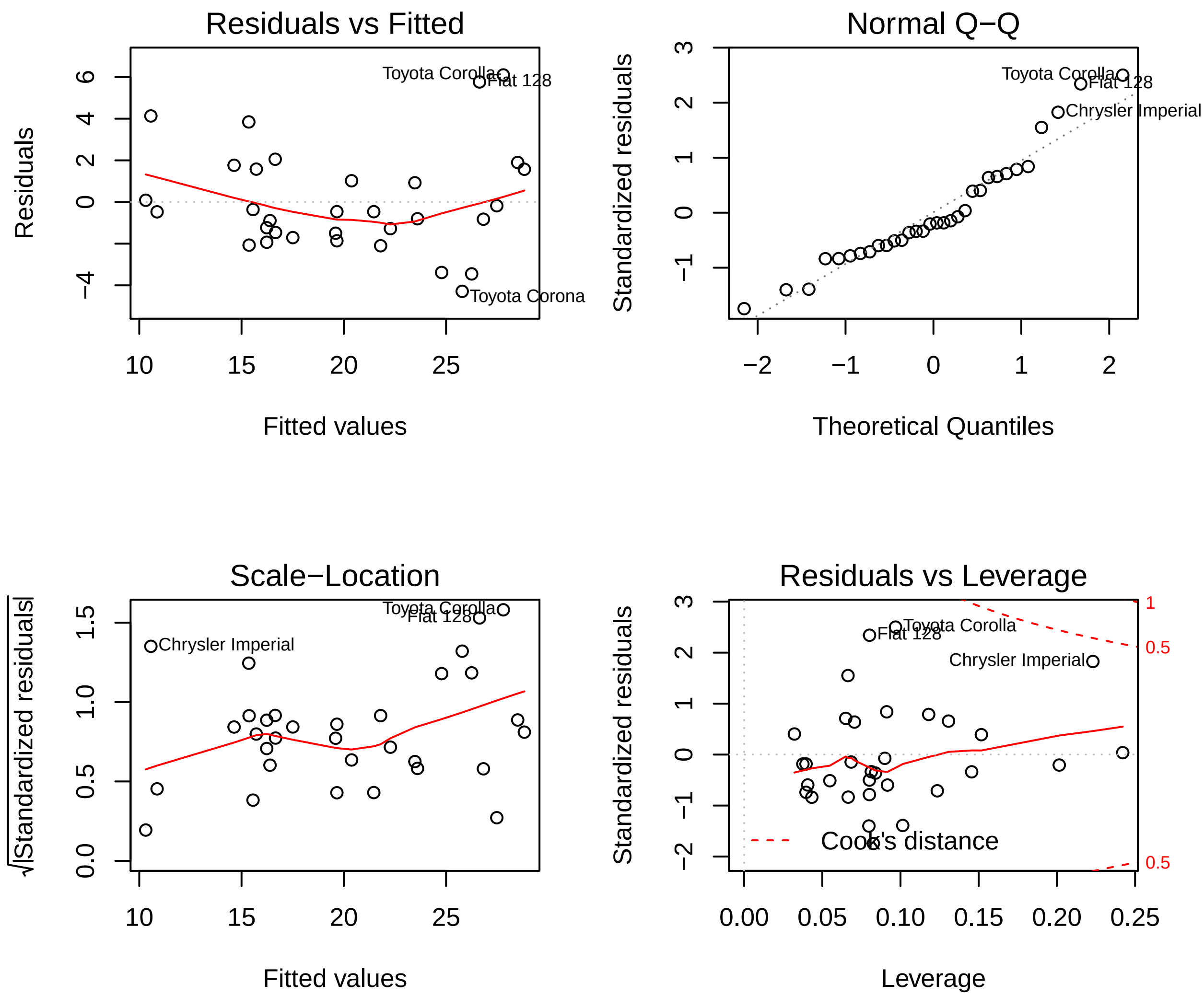

Figure 5: Standard diagnostics plots for unconstrained best model.