Motor Trend Cars Analysis

S. Kairs

Executive Summary

In this report we investigate the relationship between transmission type and fuel efficiency, as reported in the 1974 edition of Motor Trend US magazine. The analysis is prepared as part of the Regression Models course, offered by Johns Hopkins University on Coursera.

We will show that a manual transmission is more fuel efficient, and that transmission type alone is not a sufficient predictor of fuel efficiency. When other variables–such as weight, number of cylinders and horsepower–are added, the fitted model is more accurate. **Adjusted for these additional variables, a manual transmission will increase mpg by a factor of 1.8 over an automatic.**

Loading the dataset

The `mtcars` dataset is included in the `datasets` package of R and consists of fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). We load in the data set, and convert the following variables to factors: `cyl`, `vs`, `am`, `gear`, `carb`.

Exploratory Data Analysis

We first plot the pair-wise relationships between all of the variables. (Figure 1, Appendix.) We are principally concerned with the influence other variables have on fuel efficiency, or `mpg`, so we focus on the sub-plots for these interactions. Note that `cyl`, `disp`, `hp`, `drat`, `wt`, `vs` and `am` seem to have a strong correlation with mpg.

Next, we plot the `mpg` as the response and `am` as the predictor. (Figure 2, Appendix.) The resulting boxplot shows that MPG ratings for cars with manual transmissions are higher than for cars with automatic transmissions.

A Welch Two Sample t-test shows this is statistically significant, with a p-value of 0.0014. However, the r-squared value for a linear fit of this relationship, 0.3598, indicates that this model explains only 35.98% of the variance. **Transmission type alone is not a particularly good predictor of fuel efficiency.**

Regression Analysis

Multivariable Linear Regression and Model Selection

To generate a multivariable regression model, we build an initial model with `mpg` as the outcome and all other variables as predictors. Then we perform step-wise model selection by using the `step` function to build and evaluate many models. The code is shown below, but the output is suppressed for brevity.

```r
fit0 <- lm(mpg ~ ., mtcars)
fit1 <- step(fit0, direction = "both")
```

The best model includes predictor varibles `cyl`, `hp`, `wt`, and `am`. The r-squared value for this adjusted model is 0.8401. The improved model explains 84.01% of the variance in the data set. See the summary below:

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -3.939 -1.256 -0.401  1.125  5.051
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  33.7083     2.6049   12.94  7.7e-13 ***
## cyl6         -3.0313     1.4073   -2.15   0.0407 *
## cyl8         -2.1637     2.2843   -0.95   0.3523
## hp           -0.0321     0.0137   -2.35   0.0269 *
## wt           -2.4968     0.8856   -2.82   0.0091 **
## amManual      1.8092     1.3963    1.30   0.2065
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.866,  Adjusted R-squared:  0.84
## F-statistic: 33.6 on 5 and 26 DF,  p-value: 1.51e-10
```

Using ANOVA, we compare this best fit model with the model using `am` as the only predictor variable. The p-value is highly significant and we reject the null hypothesis that the confounding variables do not contribute to the accuracy of the improved model. See below:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ cyl + hp + wt + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     26 151  4       570 24.5 1.7e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residuals and diagnostics

The residual diagnostic plots for the best fit model, `fit1`, can be found in Figure 3 of the Appendix. The random scatter of the residuals vs. fitted values points verifies the assumption of indpendence (homoskedasticity). Points in the Normal Q-Q plot fall mostly on the identity line, indicating normality. We can further test the residuals for normality using a Shapiro-Wilk test. The resulting p-value of 0.4479 indicates that we cannot reject the null hypothesis or, more plainly, that the residuals are normally distributed.

Conclusions

Based on our analysis of our best fit model, we infer the following:

1. Cars with `Manual` transmissions get 1.8x more miles per gallon compared to cars with Automatic transmission, adjusted by `hp`, `cyl`, and `wt`, with a 95% CI of -1.06 to 4.68 mpg.

2. `mpg` will decrease by 2.5 (adjusted by `hp`, `cyl`, and `am`) for every 1000 lb increase in `wt`.

3. `mpg` decreases very slightly with increase of `hp` (adjusted for `cyl`, `wt` and `am`).

4. If number of cylinders, `cyl`, increases from 4 to 6 and again to 8, mpg will decrease by 3x and 2.2x respectively (adjusted by `hp`, `wt`, and `am`).

Appendix

Figure 1. Pairs Plot for `mtcars` data set

Automatic transmission data points plotted in green; manual in blue.

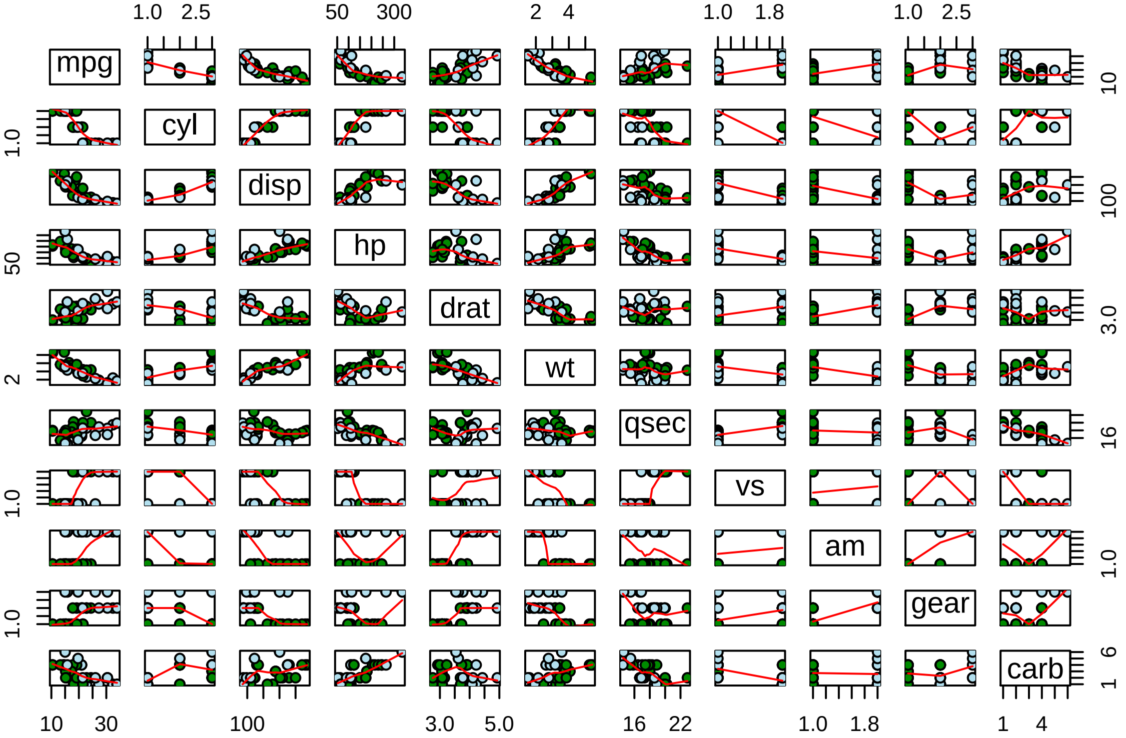


Figure 2. Boxplot of fuel efficiency vs transmission type

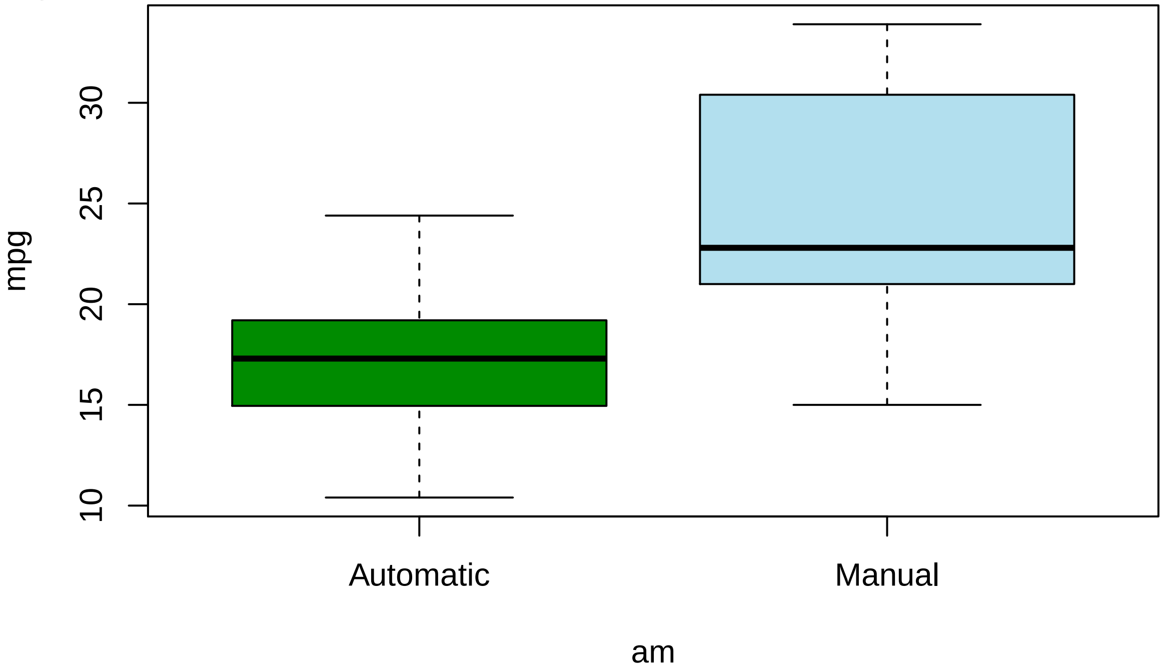
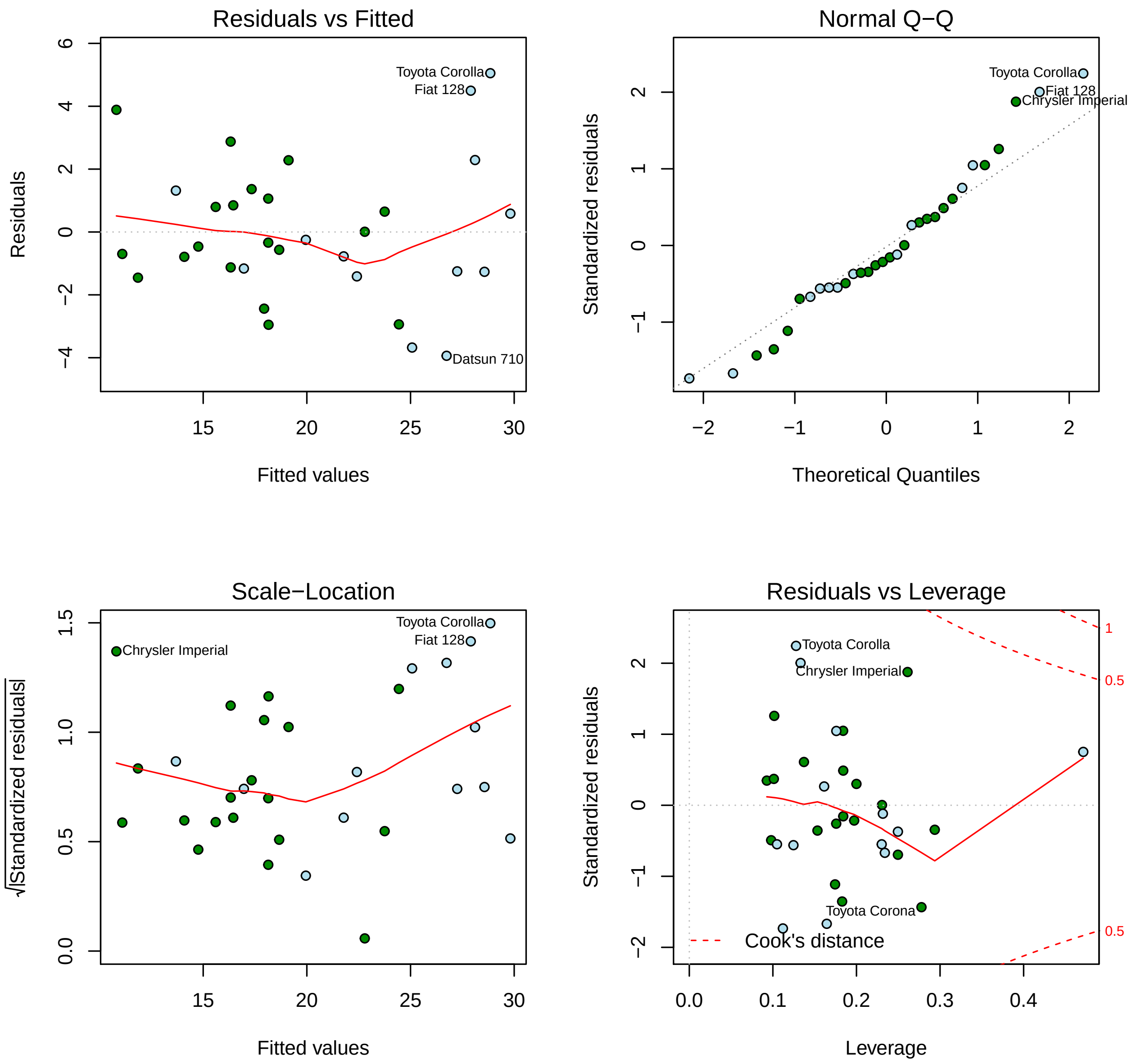


Figure 3. Residual diagnostic plots for best fit model (fit1)

Automatic transmission data points plotted in green; manual in blue.


Residuals vs Fitted
Residuals
Fitted values
6
4
2
0
−2
−4
15
20
25
30
Toyota Corolla
Fiat 128
Datsun 710
Normal Q−Q
Standardized residuals
Theoretical Quantiles
2
1
0
−1
−2
−1
0
1
2
Toyota Corolla
Fiat 128
Chrysler Imperial
Scale−Location
√|Standardized residuals|
Fitted values
1.5
1.0
0.5
0.0
15
20
25
30
Toyota Corolla
Fiat 128
Chrysler Imperial
Residuals vs Leverage
Standardized residuals
Leverage
2
1
0
−1
−2
0.0
0.1
0.2
0.3
0.4
Toyota Corolla
Chrysler Imperial
Toyota Corona
Cook's distance
1
0.5
0.5