Relationship between transmission type and miles per gallon

Synopsis

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG?"
- "Quantify the MPG difference between automatic and manual transmissions"

Exploratory Data Analysis

Data we are going to use for this document comes from the mtcars dataset. It comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models). A brief summary of data can be seen on Table 1 and Figure 1 of the Appendix.

Model selection

Let's start by fitting the simplest model we can regarding the given questions, that would be MPG as a function of transmission':

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am1            7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

The p-value of the coefficients is below our threshold of 0.05, so there's evidence to conclude that there is some relation between the type of transmission and miles per US gallon a car can achieve. However, the model has a low R-squared and so it appears to have captured little variance, underfitting the data, resulting in low generalization of their results.

Therefore, we're going to increase the number of predictors in out linear model, and test their performance with an ANOVA test.

To do that, let's select the new features using a correlation matrix. Recall that multicollinearity is a potential problem on regression analysis. Therefore, we're going to fit a second model with features that have low correlation (according to the *Figure 2* of the Appendix).

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 45.618 1.55e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA test shows as that our second model is significantly better than our first, mode simple model. Let's take a glance to this last model:

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars_dataset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## am1           2.9358     1.4109   2.081 0.046716 *
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Now, our model can explain 83% of the variability, providing a better fit. Also, looking at the plots displayed on Figure 3 of the Appendix, we can see:

- No distinct pattern in the residual vs fitted plot, thus we can infer the lack of any non-linear relationship.
- Normal distribution in the QQ plot.
- No significant outlires in the Residuals vs Leverage plot.

Conclusion

After fitting and analyzing our model, we are capable of answering the given questions:

- "Is an automatic or manual transmission better for MPG?"

Manual transmission has proven to better while considering miles per gallon consumption.

- "Quantify the MPG difference between automatic and manual transmissions".

Manual transmission delivers about 2.9 miles per gallon more than automatic transmission.

Appendix

Table 1: Data summary

Name	mtcars_dataset
Number of rows	32
Number of columns	11
Column type frequency:	
factor	2
numeric	9
Group variables	None

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
vs	0	1	FALSE	2	0: 18, 1: 14
am	0	1	FALSE	2	0: 19, 1: 13

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
mpg	0	1	20.09	6.03	10.40	15.43	19.20	22.80	33.90
cyl	0	1	6.19	1.79	4.00	4.00	6.00	8.00	8.00
disp	0	1	230.72	123.94	71.10	120.83	196.30	326.00	472.00
hp	0	1	146.69	68.56	52.00	96.50	123.00	180.00	335.00
drat	0	1	3.60	0.53	2.76	3.08	3.70	3.92	4.93
wt	0	1	3.22	0.98	1.51	2.58	3.33	3.61	5.42
qsec	0	1	17.85	1.79	14.50	16.89	17.71	18.90	22.90
gear	0	1	3.69	0.74	3.00	3.00	4.00	4.00	5.00
carb	0	1	2.81	1.62	1.00	2.00	2.00	4.00	8.00

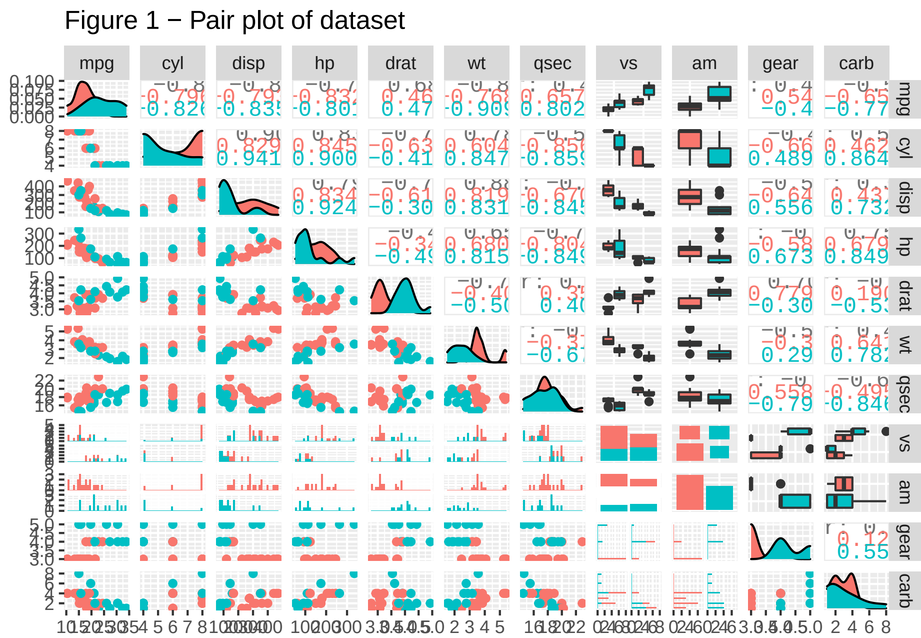


Figure 1: Pairs plot of dataset

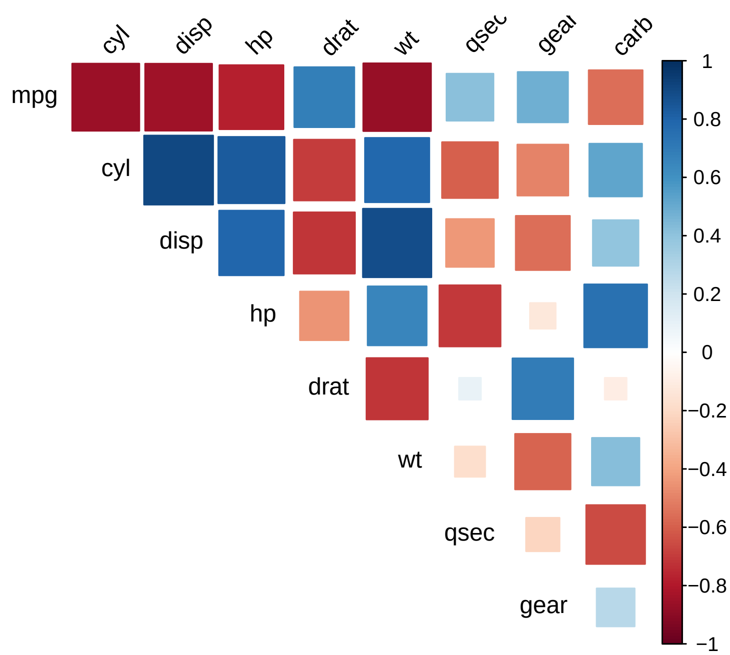
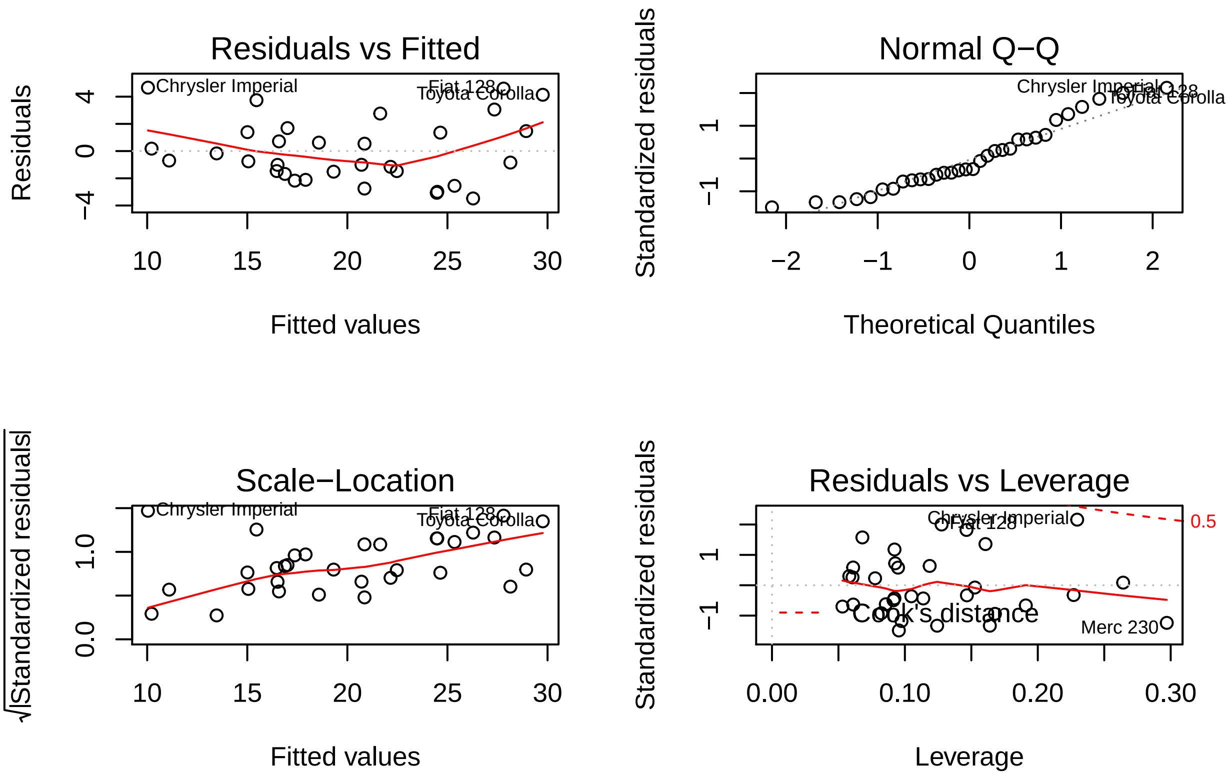


Figure 2: Correlation Matrix


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Leverage
Cook's distance
Merc 230
0.5