MPG vs Transmission Type in mtcars Dataset

Executive Summary

This is a brief exercise looking MPG (Miles/(US) gallon) and Transmission in the mtcars dataset to explore whether 1) an automatic or manual transmission better for MPG?; and to 2) quantify the MPG difference between automatic & manual transmissions.

Our final model shows that having a manual or automatic transmission doesn't have a statistically significant association with mpg once other car attributes like number of cylinders, weight, and horsepower are controlled for. While the mean mpg for automatic cars is considerably lower than manual in the data (17.2 vs 24.4), the relationship seems confounded by these variables.

The final model suggests that higher 'cyl', 'wt', and 'hp' are all significantly associated with lower MPG.

The Data

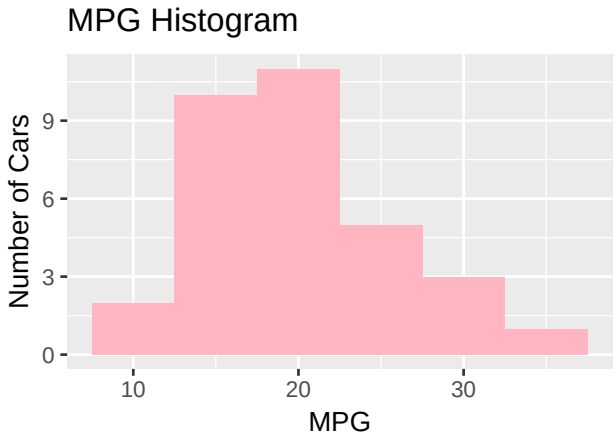
The data has 32 observations on 11 variables, 4 of which appear to be factor variables.

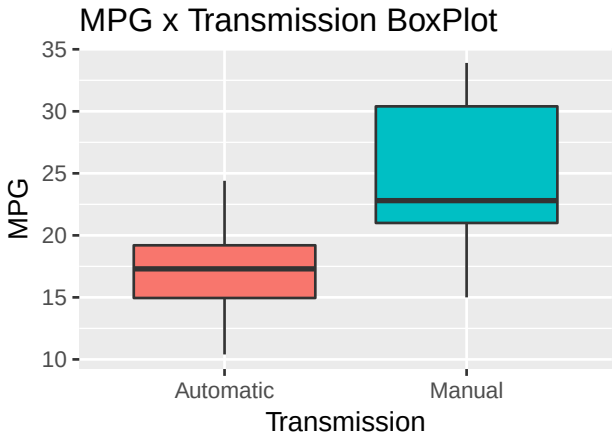
```
data(mtcars)
?mtcars
```

Having read the data documentation, we will transform factor variables from numeric using dplyr:

```
library(dplyr)
data <- mtcars %>%
  mutate(am = factor(am)) %>%
  mutate(cyl = factor(cyl)) %>%
  mutate(vs = factor(vs)) %>%
  mutate(gear = factor(gear))
```

The mean MPG for all 32 cars in the dataset is 20.1, but a boxplot suggests automatic transmissions have lower MPG (mean of 17.2 vs 24.4).





Looking at Fig.1 & 2 in the Appendix, MPG is highly correlated with 'cyl', as well as quite a few of the numeric variables, especially 'wt', 'disp', and 'hp' - these will have to be considered in the model.

Model Selection

Since MPG is a continuous variable, it will be modelled as the outcome using a multiple linear model. We will start by using Transmission (am) as the only regressor, and then create nested models to control for other variables. We will start by adding 'cyl', as there seems to be a strong correlation with mpg, and then add numeric variables with a correlation coefficient > 0.7 with MPG in descending order. To avoid issues of multicollinearity, 'disp' will be skipped, as it's highly correlated with 'wt'.

Models will be compared using ANOVA, which will help determine whether additional variables are necessary.

```
lm1 <- lm(mpg ~ am, data)
lm2 <- lm(mpg ~ am + cyl, data)
lm3 <- lm(mpg ~ am + cyl + wt, data)
lm4 <- lm(mpg ~ am + cyl + wt + hp, data)
```

The outputs from ANOVA compare a reduced model step-wise to a more full model (lm4). The outputs suggest there is good reason to use model 'lm4' over the others.

ANOVA: lm1, lm2, lm3, lm4

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	30	720.90				
2	28	264.50	2	456.40	39.29	0.0000
3	27	182.97	1	81.53	14.04	0.0009
4	26	151.03	1	31.94	5.50	0.0269

Residual/diagnostic plots for 'lm4' in Fig.4 in the Appendix show we might have some issues with outliers (20, 18, 3, 17), so we'll re-do the model without these points:

```
lm4_v2 <- lm(mpg ~ am + cyl + wt + hp, data[-c(20, 18, 3,17),])
```

This new model's residual plots in Fig.5 of the Appendix look acceptable now.

Model interpretation:

- **Intercept**: *This is the mpg when all variables = 0 (in the case of Transmission (am), this is automatic).*
- **Transmission (am)**: *Manual or Automatic is not significant once we control for cyl (compare outputs from model 1 to model 2 in Fig3 in Appendix).*
- **Number of Cylinters (cyl)**: *Higher cyl is associated with lower mpg at a statistically signiant level. 6 & 8 cylinders (cyl6 & cyl8) are associated with a -2.3 reduction in mpg compared to 4.*
- **Weight (wt)**: *Higher weight is associated with lower mpg at a statistically significant level. A 1,000 lbs increase in wt is associated with a -3.2 reduction in mpg.*
- **Gross horsepower (hp)**: *Higher horsepower is associated with lower mpg at a statistically significant level. A 1-unit increase in hp is associated with a -0.02 reduction in mpg.*
- The model has R-squared of 0.89, which explains 89% of the variability in mpg.

MPG model outputs: lm4_v2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.7201	2.0483	16.95	0.0000
am1	0.6021	1.0567	0.57	0.5746
cyl6	-2.3333	1.0446	-2.23	0.0360
cyl8	-2.3019	1.6609	-1.39	0.1796
wt	-3.2400	0.7056	-4.59	0.0001
hp	-0.0229	0.0100	-2.28	0.0326

APPENDIX

Fig.1: Scatterplot Matrix: Scatterplots for Pairs of Variables in MtCars Data

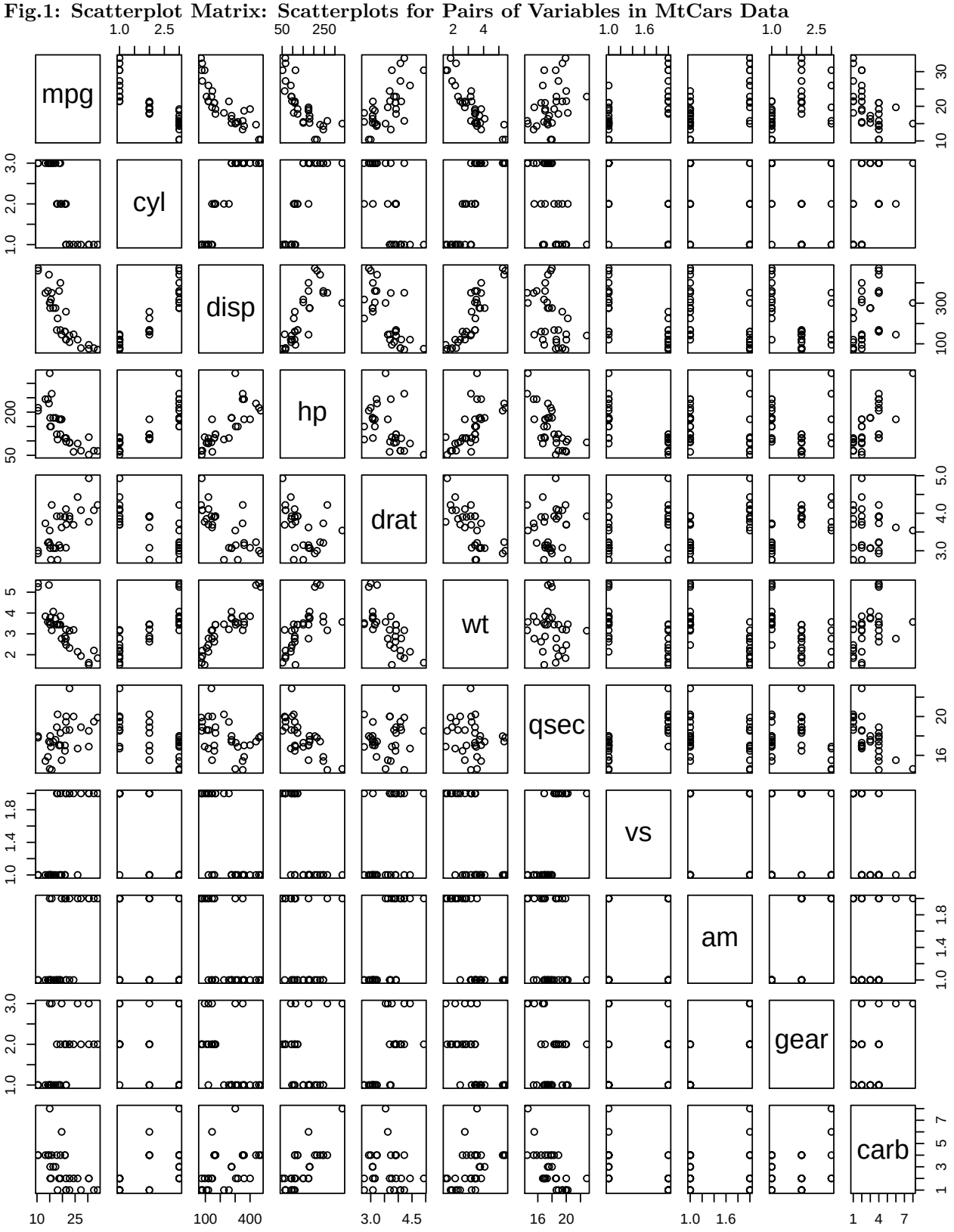


Fig.2: Correlation of MtCars Numeric Variables

	mpg	disp	hp	drat	wt	qsec	carb
mpg	1.00	-0.85	-0.78	0.68	-0.87	0.42	-0.55
disp	-0.85	1.00	0.79	-0.71	0.89	-0.43	0.39
hp	-0.78	0.79	1.00	-0.45	0.66	-0.71	0.75
drat	0.68	-0.71	-0.45	1.00	-0.71	0.09	-0.09
wt	-0.87	0.89	0.66	-0.71	1.00	-0.17	0.43
qsec	0.42	-0.43	-0.71	0.09	-0.17	1.00	-0.66
carb	-0.55	0.39	0.75	-0.09	0.43	-0.66	1.00

Fig.3: Model Outputs

Model 1:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.1474	1.1246	15.25	0.0000
am1	7.2449	1.7644	4.11	0.0003

Model 2:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	24.8019	1.3226	18.75	0.0000
am1	2.5600	1.2976	1.97	0.0585
cyl6	-6.1561	1.5357	-4.01	0.0004
cyl8	-10.0676	1.4521	-6.93	0.0000

Model 3:

Model 4:

Fig.4: lm4/Model4 Residual/Diagnostic Plots

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.7536	2.8135	12.00	0.0000
am1	0.1501	1.3002	0.12	0.9089
cyl6	-4.2573	1.4112	-3.02	0.0055
cyl8	-6.0791	1.6837	-3.61	0.0012
wt	-3.1496	0.9080	-3.47	0.0018

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.7083	2.6049	12.94	0.0000
am1	1.8092	1.3963	1.30	0.2065
cyl6	-3.0313	1.4073	-2.15	0.0407
cyl8	-2.1637	2.2843	-0.95	0.3523
wt	-2.4968	0.8856	-2.82	0.0091
hp	-0.0321	0.0137	-2.35	0.0269

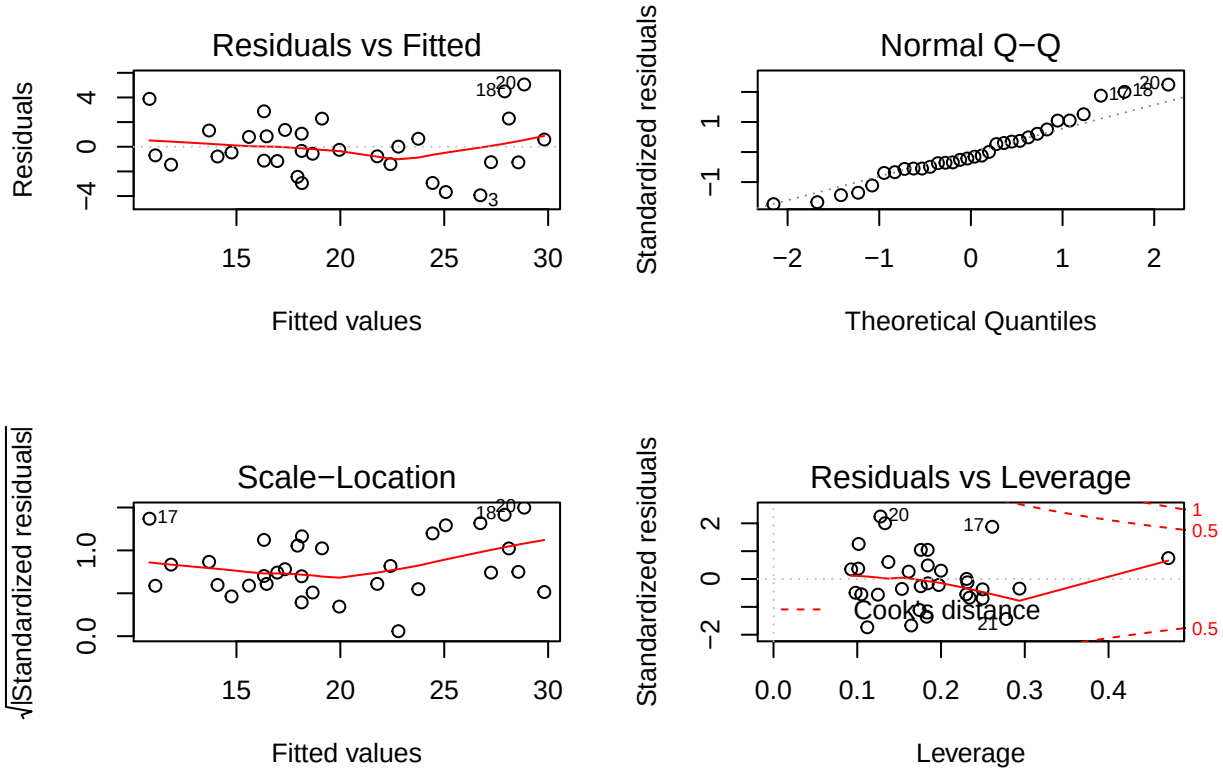


Fig.5: lm4 without outliers Residual/Diagnostic Plots

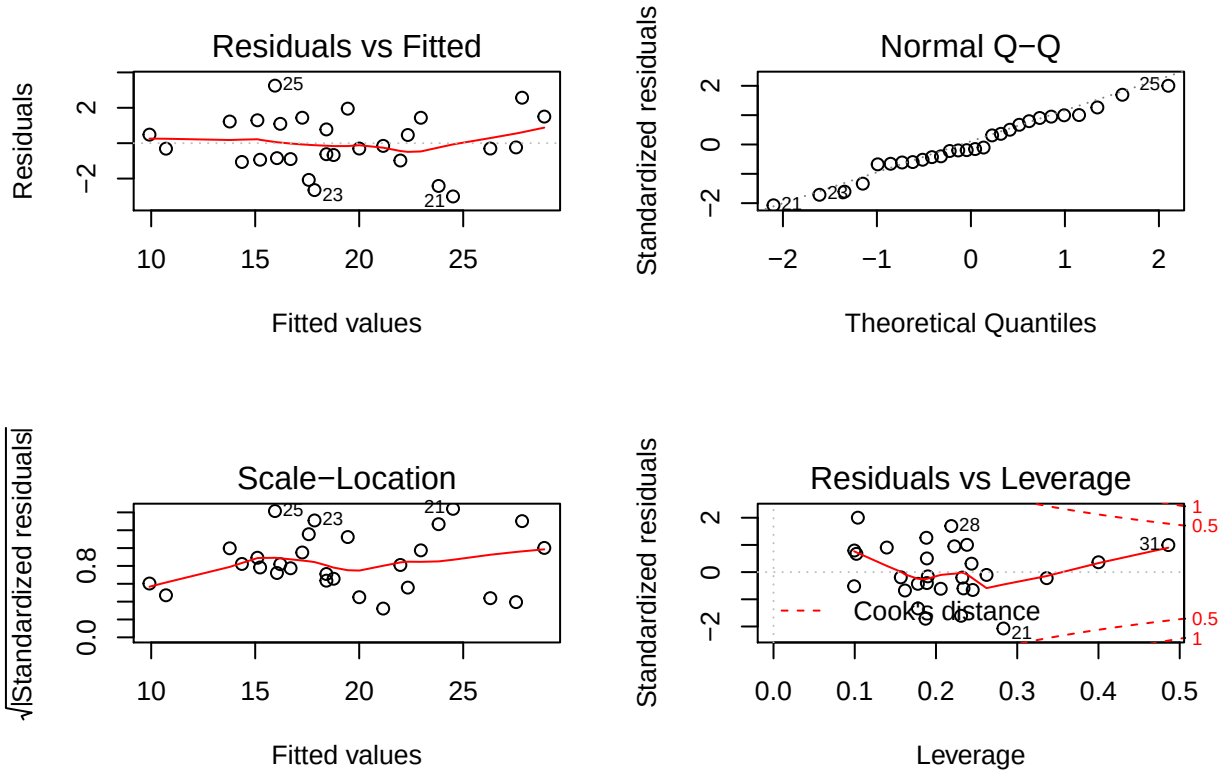


Fig.6: Relationship between Trasmission & MPG & Other Variables

Transmission vs Number of Cylinders:

	Cylinders	Automatic	Manual
1	4	3	8
2	6	4	3
3	8	12	2

Transmission vs Weight/Horsepower: