Analysis of the effect of type of transmission (manual / automatic) on mpg (miles per gallon obtained) for passenger vehicles

This analysis was performed on the dataset 'mtcars' with the aim of identifying the simplest, least biased, and most accurate linear relationship between 'mpg' (miles per gallon) as the outcome, and the other variables as predictors. We find that 'mpg' has a linear relationship with 'qsec' (1/4 mile time), 'wt' (weight of the vehicle), and 'am' (type of transmission where 1: manual, 0: automatic).

On an average, for a vehicle of a given qsec specification, and a given weight, changing the transmission type from automatic to manual will increase the mpg by 4.3 miles per gallon (with a 95 % confidence interval of +- 2.04 miles per gallon)

Exploratory data analysis

As the starting point, we use theoretical background: miles obtained per gallon for a vehicle depend mainly on (1) the size / weight of the vehicle (bigger vehicle = lower mileage) and (2) the power rating of the engine (more powerful = lower mileage). We find the correlation (and use scatter plots ***fig. 1***) between mpg, and each of the following regressors: weight(-ve correlation), no of cylinders (-ve), displacement(-ve), horsepower (-ve), rear axle ratio (+ve), 1/4 mile time (+ve), transmission type (manual = higher mpg).

Selecting a linear regression model

Based on the exploratory analysis, we start by fitting a multi variable linear model as follows:

fit_1 <- lm(mpg ~ hp + wt + am - 1, data = mtcars)

In the summary of the model obtained, we aim for two things:

1. Ensuring that the t-value of each regressor's coefficient is statistically significant (i.e. it exceeds the corresponding p-value)
2. Obtaining the largest possible adjusted R-squared (closer to 1) using the fewest possible regressors

fit_1 satisfies condition #1 above, and adjusted R-squared is as follows:

Adjusted R-squared: 0.7036

These are the other models attempted:

fit_2 <- lm(mpg ~ wt + am - 1, data = mtcars) Adjusted R-squared: 0.87

fit_3 <- lm(mpg ~ displ + wt + am - 1, data = mtcars) Adjusted R-squared: 0.9014

fit_4 <- lm(mpg ~ hp + disp + wt + am - 1, data = mtcars) Adjusted R-squared: 0.9057

fit_5 <- lm(formula = mpg ~ hp + cyl + wt + am - 1, data = mtcars) Adjusted R-squared: 0.9167

fit_6 <- lm(mpg ~ hp + qsec + wt + am - 1, data = mtcars) Adjusted R-squared: 0.9853 fit_7 <- lm(mpg ~ qsec + wt + am - 1, data = mtcars) Adjusted R-squared: 0.9858

fit_8 <- lm(mpg ~ qsec + am -1, data = mtcars) Adjusted R-squared: 0.9656

We further examine fit_7 and fit_8 by the following criteria:

1. The scatter plot of the residuals against the model's fitted value should be patternless

2. The respective scatter plots of the residuals against each of the left-out independent variables should be patternless

In case of any visual uncertainty as to the existence of a pattern in the residuals, we attempt to fit a linear model between the residuals and the respective variable, and examine the coefficients obtained for statistical significance.

In fit_8, we find that the residuals have a visible ***fig. 2*** and statistically significant relationship with the left-out independent variable wt (weight), hence wt must be added back to the model, which takes us to the model fit_7

fit_7 <- lm(mpg ~ qsec + wt + am - 1, data = mtcars) passes all the criteria 888fig. 3 & 4***, and we select it as our linear model

```
fit_7 <- lm(mpg ~ qsec + wt + am - 1, data = mtcars)
summary(fit_7)

##
## Call:
## lm(formula = mpg ~ qsec + wt + am - 1, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8820 -1.5401 -0.4246  1.6623  4.1711
##
## Coefficients:
##      Estimate Std. Error t value Pr(>|t|)
## qsec   1.5998     0.1021  15.665 1.09e-15 ***
## wt    -3.1855     0.4828  -6.598 3.13e-07 ***
## am     4.2995     1.0241   4.198 0.000233 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.497 on 29 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.9858
## F-statistic:   741 on 3 and 29 DF,  p-value: < 2.2e-16
```

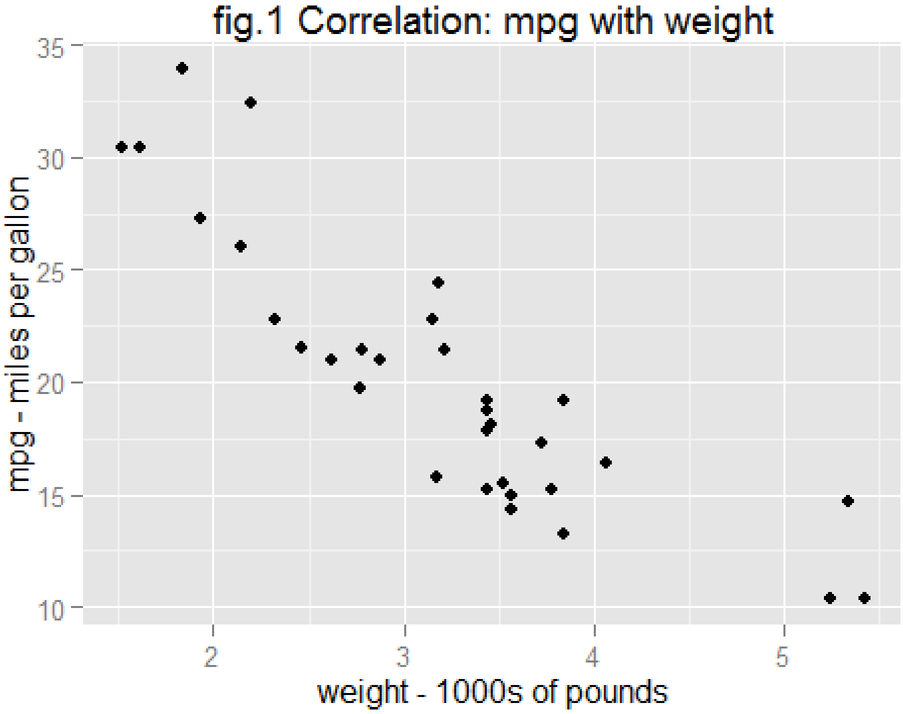
As seen above, the coefficient of ‘am’ in the linear model is 4.2995 (approx. 4.3). This is interpreted as the change in the outcome (mpg) per unit change in the regressor ‘am’, while the other two regressors (wt, qsec) are kept constant.

This is the end of the main text of the analysis.

Appendix

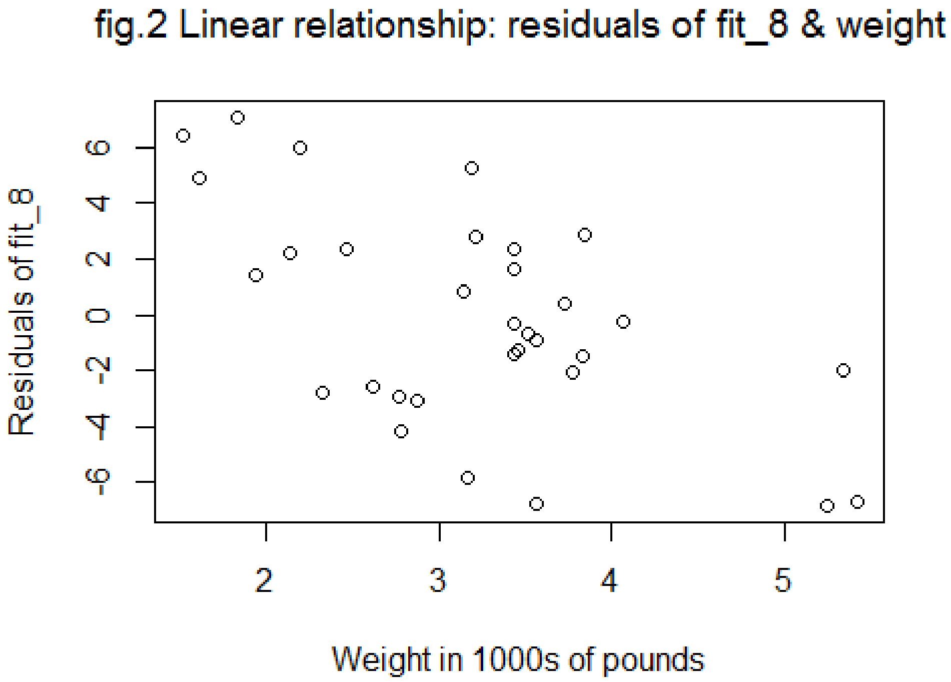
Example of Correlation scatter plot between theoretically relevant variables

```
library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.1.3

g <- ggplot(mtcars, aes(y = mpg, x = wt))
g <- g + geom_point()
g <- g + xlab("weight - 1000s of pounds")
g <- g + ylab("mpg - miles per gallon")
g<- g + ggtitle("fig.1 Correlation: mpg with weight")
g
```



Linear relationship between resid(fit_8) and mtcars$wt

```
fit_8 <- lm(mpg ~ qsec + am -1, data = mtcars)
plot(mtcars$wt, resid(fit_8), xlab = "Weight in 1000s of pounds", ylab =
"Residuals of fit_8", main = list("fig.2 Linear relationship: residuals of
fit_8 & weight", font = 1))
```

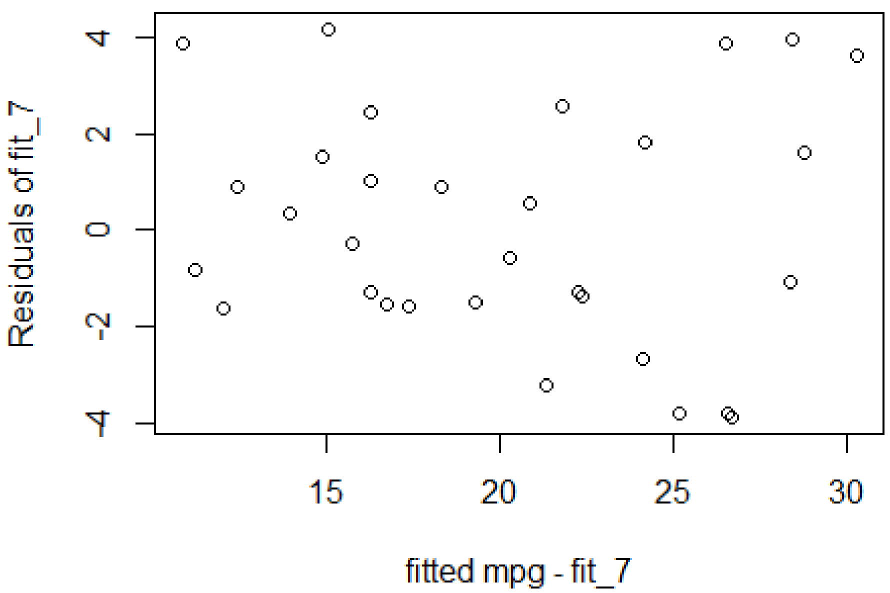


```
summary(lm(I(resid(fit_8) ~ mtcars$wt)))

##
## Call:
## lm(formula = I(resid(fit_8) ~ mtcars$wt))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7953 -1.5759  0.1619  2.4286  5.3360
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   7.2335     1.9315   3.745 0.000765 ***
## mtcars$wt    -2.2996     0.5751  -3.998 0.000384 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.133 on 30 degrees of freedom
## Multiple R-squared:  0.3476, Adjusted R-squared:  0.3259
## F-statistic: 15.99 on 1 and 30 DF,  p-value: 0.0003835
```

No relationship between resid(fit_7) and predict(fit_7)

```
plot(predict(fit_7), resid(fit_7), xlab = "fitted mpg - fit_7", ylab =
"Residuals of fit_7", main = list("fig. 3 No pattern: residuals & fitted
points of fit_7", font = 1))
```

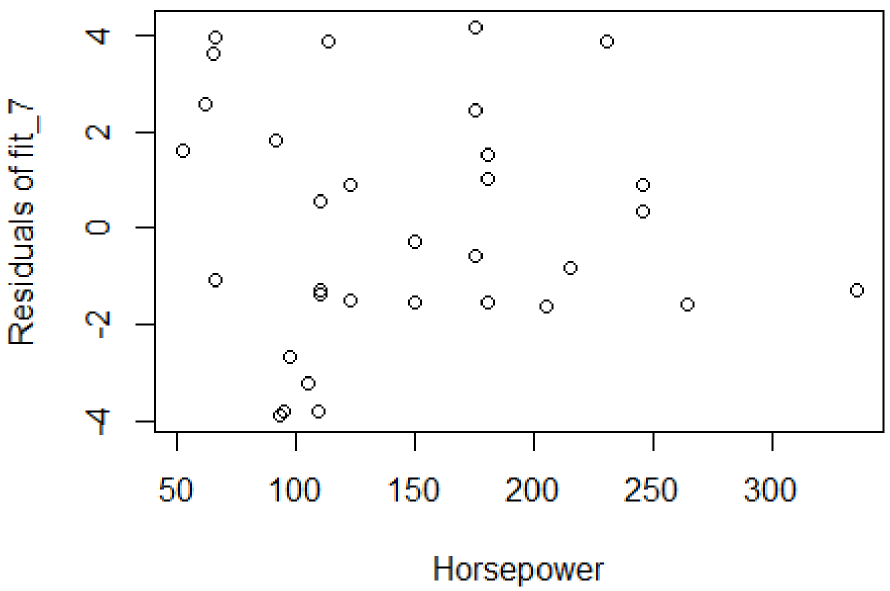
fig. 3 No pattern: residuals & fitted points of fit_7



Example: no relationship between resid(fit_7) and a left-out independent variable

```
plot(mtcars$hp, resid(fit_7), xlab = "Horsepower", ylab = "Residuals of
fit_7", main = list("fig. 4 No pattern: residuals of fit_7 & horsepower",
font = 1))
```

fig. 4 No pattern: residuals of fit_7 & horsepower


Residuals of fit_7
Horsepower