Course Project: Regression Models

Executive Summary

This study is an analysis of the mtcars dataset in which I have investigated the effect that transmission type has on fuel efficiency. I have compared a series of regression models to develop a comprehensive model that fits the observations with reasonable precision. The resultant model shows efficiency (mpg) vs weight (wt), number of cylinders (cyl) and transmission type (am): mpg ~ wt + cyl + am. In this fitted model, manual transmissions are associated with an increase of 0.15 mpg, all other factors being equal. This small change is not statistically significant at $p < 0.05$, leading me to conclue that there is no evidence that transmission type per se has an effect on fuel efficiency.

Loading data and exploratory analysis

Data were loaded into R with the following commands.

```
data("mtcars")
library(ggplot2); require(GGally)
cars <- mtcars
cars$trans <- ifelse(cars$am == 0, "automatic", "manual")
```

A simple analysis of fuel efficiency for manual versus automatic transmission shows that if we ignore any other factor, cars with manual transmissions appear to have greater fuel efficiency (Fig. 1). Yet there may be other confounding factors at play here.

Results

Correlation analysis

To explore what factors might have the greatest effect on fuel efficiency, I performed pair-wise correlation analysis with mpg versus all other factors in the data set:

```
cor(cars$mpg, cars[,-12])
```

```
##      mpg       cyl       disp         hp      drat         wt     qsec
## [1,]   1 -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##             vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

As we can see, the greatest correlation is between mpg and wt. cyl and disp also appear to have a strong correlation with mpg, yet transmission type (am) is among the lowest correlations.

Models

To develop a working model, I first produced a model mpg ~ 1 and used the add1 command to see which factor would have the greatest impact on improving the fit.

```
fit1 <- lm(mpg ~ 1, data = cars)
add1(fit1, scope =cars[,-12], test = "F")
```

This analysis (see appendix for output) shows that wt is the variable that most strongly affects mpg. I added wt to the model and performed the analysis again. I repeated the process until there were no longer missing variables that would give a statisical improvement to the model. With this method, the variable cyl was added as a factor. To address the question of the study, the am variable was also added, although ANOVA shows that this addition makes no improvement to the model (appendix). The final model was called fit4.

```
fit4 <- lm(mpg ~ wt + as.factor(cyl) + am, data = cars)
summary(fit4)$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     33.7535920  2.8134831 11.9970836 2.495549e-12
## wt              -3.1495978  0.9080495 -3.4685309 1.770987e-03
## as.factor(cyl)6 -4.2573185  1.4112394 -3.0167231 5.514697e-03
## as.factor(cyl)8 -6.0791189  1.6837131 -3.6105432 1.227964e-03
## am               0.1501031  1.3002231  0.1154441 9.089474e-01
```

The addition of the binary am variable gives a coefficient that affects the y intercept of the model. This is found under the column heading Estimate for the am variable (fifth row). The value 0.15 indicates that, all other factors being equal, manual transmissions (am = 1) are correlated with a 0.15 mpg increase in fuel efficiency. Note the high p value for this variable and low T statistic, indicating that we can be 90% confident that transmission type does not affect fuel efficiency.

Diagnostics

To analyse the fit of the model, I put it through several standard diagnostic tests (Fig. 2). Most notably, the residual plot (top left panel), shows no obvious pattern across the distribution.

Figure 3 shows a plot of the data. This includes regression lines and 95% confidence intervals for each of the three factors in the cyl variable. I have also drawn regression lines for the linear model mpg ~ wt with (blue) and without (red) the inclusion of the am variable. As we can see, inclusion of am has almost no effect on the fit of the model.

Conclusions

In this study, I created a multivariable linear regression model to describe the mtcars data. I found that manual transmissions are correlated with a 0.15 mpg increase in fuel efficiency, but that this is not statistically significant.

APPENDIX

Figure 1

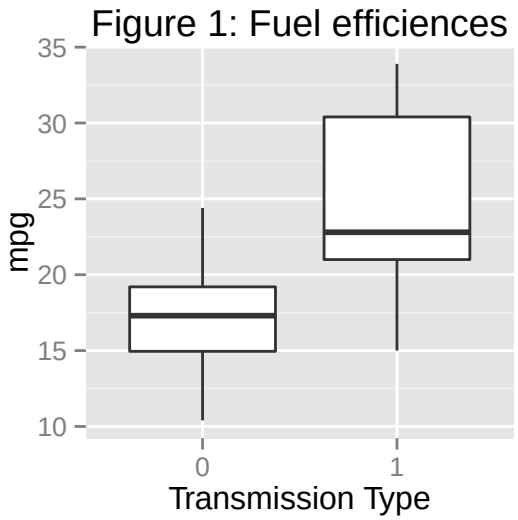

Figure 2: Output of add1 command on model fit1

```
fit1 <- lm(mpg ~ 1, data = cars)
add1(fit1, scope =cars[,-12], test = "F")
```

```
## Single term additions
##
## Model:
## mpg ~ 1
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>              1126.05 115.943
## cyl     1    817.71  308.33  76.494 79.5610 6.113e-10 ***
## disp    1    808.89  317.16  77.397 76.5127 9.380e-10 ***
## hp      1    678.37  447.67  88.427 45.4598 1.788e-07 ***
## drat    1    522.48  603.57  97.988 25.9696 1.776e-05 ***
## wt      1    847.73  278.32  73.217 91.3753 1.294e-10 ***
## qsec    1    197.39  928.66 111.776  6.3767  0.017082 *
## vs      1    496.53  629.52  99.335 23.6622 3.416e-05 ***
## am      1    405.15  720.90 103.672 16.8603  0.000285 ***
## gear    1    259.75  866.30 109.552  8.9951  0.005401 **
## carb    1    341.78  784.27 106.369 13.0736  0.001084 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: ANOVA on models

An ANOVA test comparing the four models confirms that the addition of the am variable adds no precision to the model, indicating that there is no evidence that transmission type affects mpg, all other factors being equal.

```r
anova(fit1, fit2, fit3f, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ 1
## Model 2: mpg ~ wt
## Model 3: mpg ~ wt + as.factor(cyl)
## Model 4: mpg ~ wt + as.factor(cyl) + am
##   Res.Df     RSS Df Sum of Sq        F   Pr(>F)
## 1     31 1126.05
## 2     30  278.32  1    847.73 125.0959 1.22e-11 ***
## 3     28  183.06  2     95.26   7.0288 0.003488 **
## 4     27  182.97  1      0.09   0.0133 0.908947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 4: Diagnostic plots on fit4 model

```r
par(mfrow = c(2,2)); plot(fit4, main = "Figure 4")
```

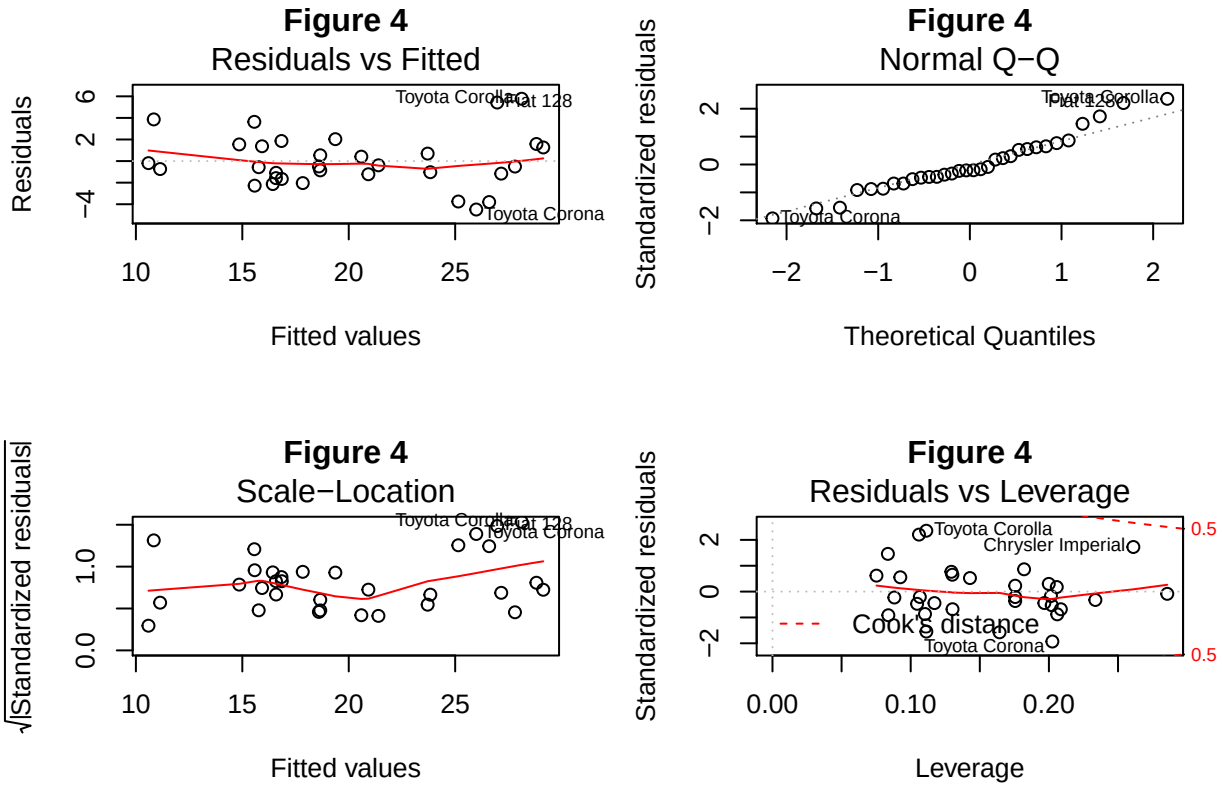

Figure 5: plotting the data points

To see a summary of the data, I plotted mpg versus wt and colour coded for the cyl variable. I have overlaid a regression line for each cyl factor with 95% confidence intervals for regression fit in each. The overlaid red and blue lines are the regression fit for mgp ~ wt + ab, showing the effect of transmission type.