Regression Model Project

ahmed Bakhouz

Saturday, June 20, 2015

Executive Summary

This paper explores the relationship between miles-per-gallon (MPG) and other variables in the mtcars data set. In particular, the analysis attempts to determine whether an automatic or manual transmission is better for MPG, and quantifies the MPG difference.

1. The data Description

The data set was extracted from the 1974 edition of Motor Trend US Magazine and it deals with 1973 - 1974 models. It consists of 32 observations on 11 variables:

- `mpg`: Miles per US gallon
- `cyl`: Number of cylinders ()
- `disp`: Displacement (cubic inches)
- `hp`: Gross horsepower
- `drat`: Rear axle ratio
- `wt`: Weight (lb / 1000)
- `qsec`: 1 / 4 mile time
- `vs`: V/S
- `am`: Transmission (0 = automatic, 1 = manual)
- `gear`: Number of forward gears
- `carb`: Number of carburetors

2. Analysis

2.1 Simple Linear Regression - lm(mpg ~ am, data = mtcars)

The exploratory analysis of the data is described in Appendix. Based on the exploratory analysis, we selected three models to explore the question posed by this report:

```
data(mtcars)
n <- length(mtcars$mpg)
alpha <- 0.05
fit <- lm(mpg ~ am, data = mtcars)
coef(summary(fit))
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

The beta0 / intercept coefficient is mean MPG for cars with automatic transmissions; the beta1 / am coefficient is the mean increase in MPG for cars with manual transmissions (am = 1). The sum beta0 + beta1 is our mean MPG for cars with manual transmissions.

Using the output above, we can calculate a 95% confidence interval for beta1 (mean MPG difference) as follows:

```
pe <- coef(summary(fit))["am", "Estimate"]
se <- coef(summary(fit))["am", "Std. Error"]
tstat <- qt(1 - alpha/2, n - 2)  # n - 2 for model with intercept and slope
pe + c(-1, 1) * (se * tstat)
```

```
## [1]  3.64151 10.84837
```

The p-value of 2.850207410^{-4} for beta1 is small and the CI does not include zero, so we can reject null in favor of the alternative hypothesis that there is a significant difference in MPG between the two groups at alpha = 0.05.

2.2 Multiple Regression - lm(mpg ~ wt + qsec + am, data=mtcars)

The predictors wt (weight), qsec (1/4 mile time) and am (transmission type) were first selected in an automated fashion using the bestglm package. This set of predictors yields the highest adjusted R-squared. This result agrees with what you arrive at by following this logic: 1.Start with the predictor whose correlation with mpg is highest (wt); 2.Eliminate from the model variables that are highly correlated with wt; 3.Add the remaining predictor, qsec, which is nearly orthogonal to wt; and 4.Add our variable of interest, am, to see if it is a significant predictor.

```
# fit a model using the regressors suggested by bestglm residual plot is in
# Appendix
bestfit <- lm(mpg ~ wt + qsec + am, data = mtcars)
coef(summary(bestfit))
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

Using the output above, we can calculate a 95% confidence interval for beta3 / am as follows:

```r
pe <- coef(summary(bestfit))["am", "Estimate"]
se <- coef(summary(bestfit))["am", "Std. Error"]
tstat <- qt(1 - alpha/2, n - 2)  # n - 2 for model with intercept and slope
pe + c(-1, 1) * (se * tstat)
```

```
## [1] 0.05438576 5.81728862
```

The p-value of 0.0467155 for beta3 is small and the CI does not include zero, so we can reject null in favor of the alternative hypothesis that there is a significant difference in MPG between the two groups at alpha = 0.05.

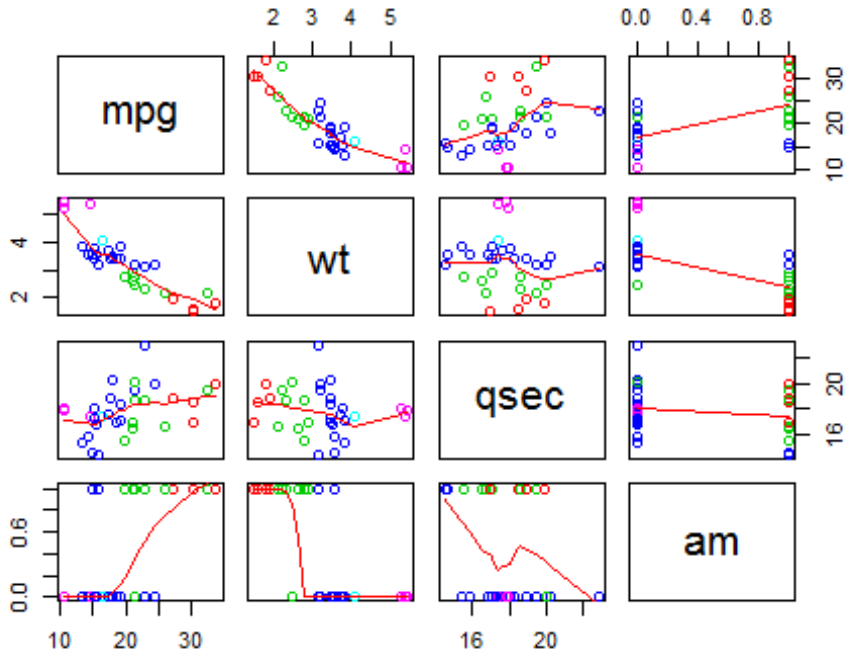
2.3 Nested Model Testing:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + qsec
## Model 3: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 278.32
## 2     29 195.46  1    82.858 13.7048 0.0009286 ***
## 3     28 169.29  1    26.178  4.3298 0.0467155 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The nested model test demonstrated in Prof. Caffo's lecture confirms that all three regressors are significant.

Appendix - Exploratory Analysis and Visualizations

2.4 Correlations

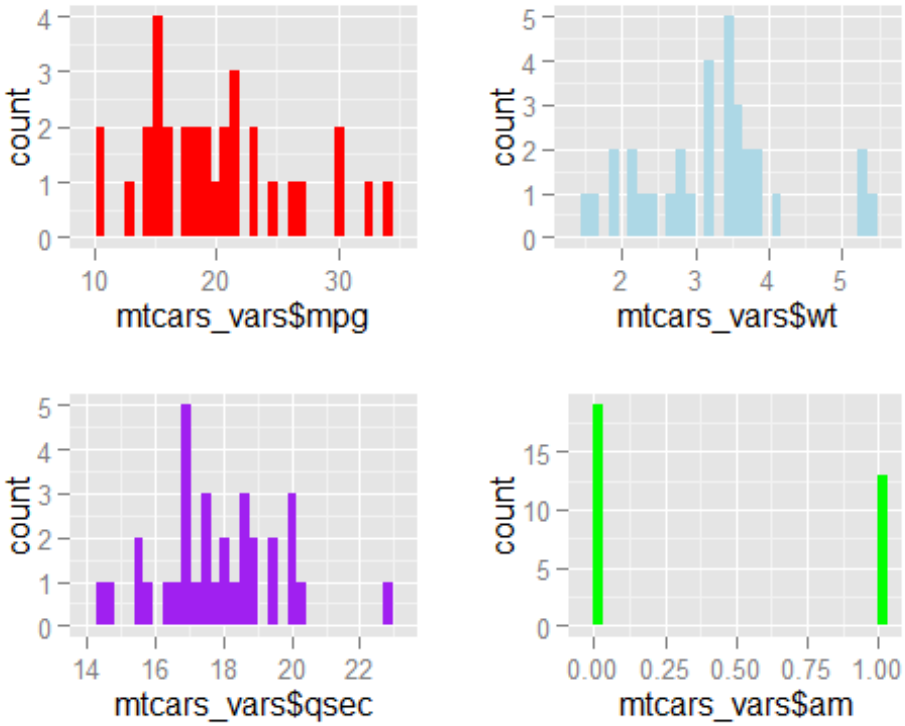



```
##              mpg         wt       qsec         am
## mpg   1.0000000 -0.8676594  0.4186840  0.5998324
## wt   -0.8676594  1.0000000 -0.1747159 -0.6924953
## qsec  0.4186840 -0.1747159  1.0000000 -0.2298609
## am    0.5998324 -0.6924953 -0.2298609  1.0000000
```

2.5 Histograms

Nothing remarkable here except perhaps in the weight / wt histogram. The Cadillac Fleetwood, Lincoln Continental and Chrysler Imperial are quite a bit heavier than other cars in the dataset.

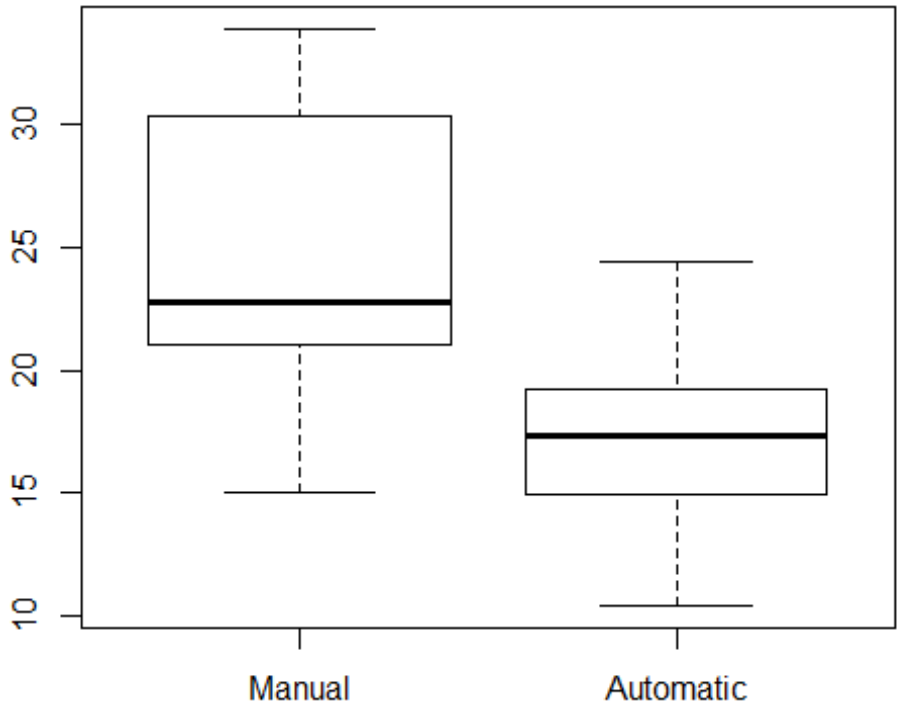
```
## Loading required package: grid
```



2.6 Homogeneity of Variance Assumption

Box plots, comparison of the standard deviations of MPG by transmission type, and Levene's test indicate that the assumption of homogeneity of variance is questionable.

Side-by-side box plots



2.7 Standard Deviation of MPG by Transmission Type

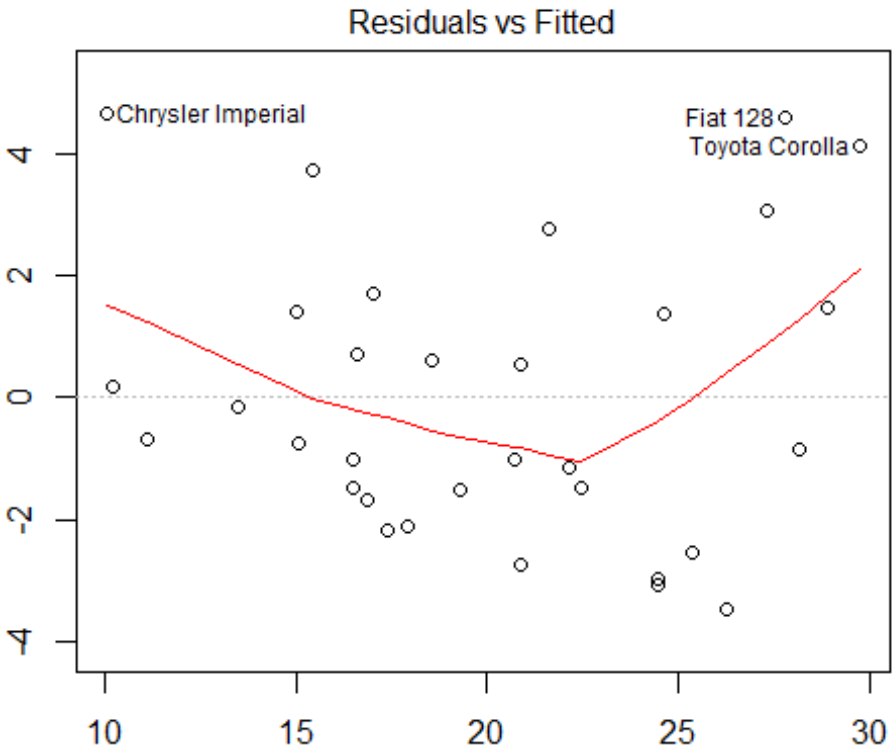
```
## mtcars_vars$am: 0
## [1] 3.833966
## ------------------------------------------------------------
## mtcars_vars$am: 1
## [1] 6.166504
```

2.8 Levene's Test for Homogeneity of Variance

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value  Pr(>F)
## group  1  4.1876 0.04957 *
##       30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.9 Residual Plot

There is a bit of a curve to the residual plot, so that it departs slightly from normality. The residuals for the Chrysler Imperial, Fiat 128, and Toyota Corolla are called out because they exert some influence on the shape of the curve.


Residuals vs Fitted
Chrysler Imperial
Fiat 128
Toyota Corolla
4
2
0
-2
-4
10
15
20
25
30