Regression Models Project

Abhilash Itharaju

July 22, 2015

Synopsis

This report answers the following questions using data obtained by Motor Trend in 1974

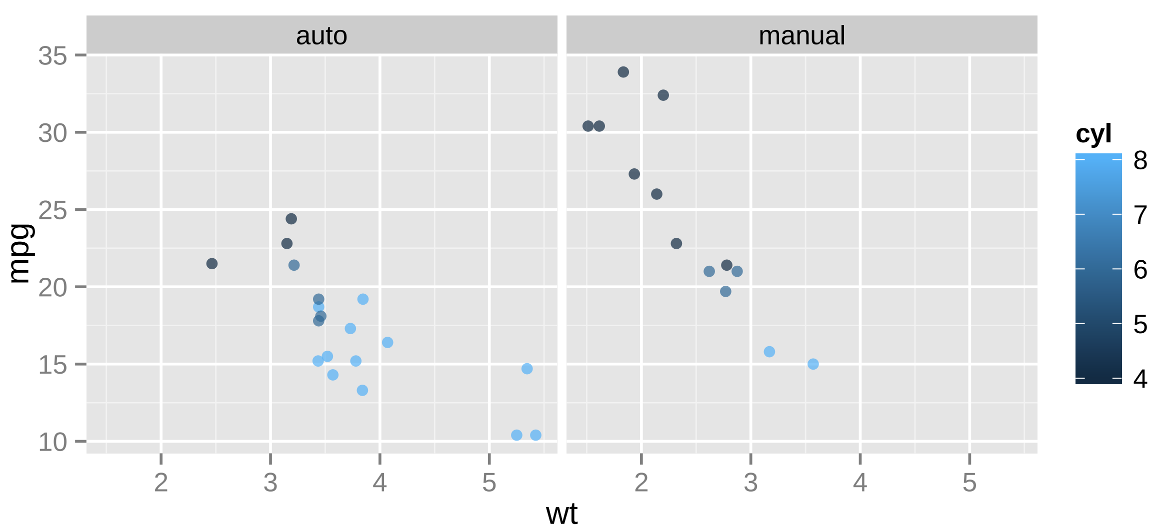
- Is an automatic or manual transmission better for MPG
- Quantify the MPG difference between automatic and manual transmissions

Exploring *mtcars* data

Let us do some data massaging and explore the data.

```
cars <- mtcars
cars$am <- as.character(cars$am)
cars$am[cars$am == '0'] <- 'auto'
cars$am[cars$am == '1'] <- 'manual'
cars$am <- as.factor(cars$am)
```

```
p <- ggplot(data=cars)
p + geom_point(aes(x=wt, y=mpg, col=cyl), size = 2, alpha=0.7) + facet_grid(. ~ am)
```



First look, it appears that manaul trasmission cars have better mpg than automatic transmission. However, the plot tells something interesting

1.Manual cars seem to start at lower weight and there are quite a few light weight cars among manual transmission cars that are lower weight than any auto car in data set

2.There a few automatic transmission cars that are heavier than any other manual transmision cars

3.Holding the weight constant, the relationship between transmission type and mpg is not apparent

Model Selection

This section explores a few models to predict mpg

Fit 1: mpg vs am

This is the simplest and direct model between mpg and am. This is literally writing the question down as a linear regression problem.

```
fit1 <- lm(mpg ~ am - 1, data = cars)
resid1 <- cars$mpg - fit1$fitted.values
summary(fit1)
```

```
##
## Call:
## lm(formula = mpg ~ am - 1, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##          Estimate Std. Error t value Pr(>|t|)
## amauto     17.147      1.125   15.25 1.13e-15 ***
## ammanual   24.392      1.360   17.94  < 2e-16 ***
## ---
## 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:  0.9487, Adjusted R-squared:  0.9452
## F-statistic: 277.2 on 2 and 30 DF,  p-value: < 2.2e-16
```

Fit 2: mpg vs am + wt

One of the observations from exploratory analysis is that auto transmission cars weight more. Do they have lesser mileage because of the extra weight? One way to know is by adjusting for weight.

```
fit2 <- lm(mpg ~ am + wt, data = cars)
resid2 <- cars$mpg - fit2$fitted.values
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5295 -2.3619 -0.1317  1.4025  6.8782
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 37.32155    3.05464  12.218 5.84e-13 ***
## ammanual    -0.02362    1.54565  -0.015    0.988
## wt          -5.35281    0.78824  -6.791 1.87e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.098 on 29 degrees of freedom
## Multiple R-squared:  0.7528, Adjusted R-squared:  0.7358
## F-statistic: 44.17 on 2 and 29 DF,  p-value: 1.579e-09
```

Fit 3: mpg vs am + wt + hp

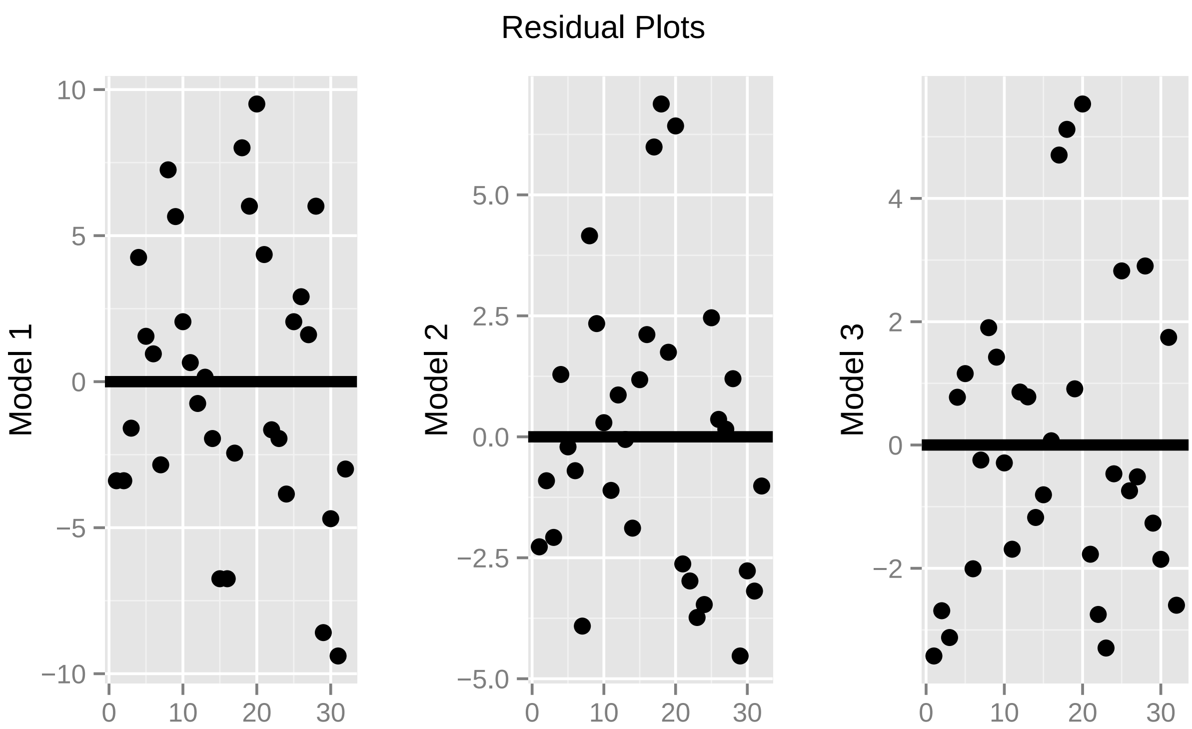
More power a engine has, more fuel it tends to draw. Hence horse power may have an effect on mpg. It is reasonable to consider horse power in the model

```
fit3 <- lm(mpg ~ am + wt + hp, data = cars)
resid3 <- cars$mpg - fit3$fitted.values
summary(fit3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## ammanual     2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

Diagnositics and Selecting a model

```
library(ggplot2)
library(grid)
library(gridExtra)
plot <- ggplot(data = cars)
p1 <- plot + geom_point(aes(x=seq_along(resid1) ,y=resid1), size = 3) + geom_hline(y=0, size = 2) + labs
p2 <- plot + geom_point(aes(x=seq_along(resid2) ,y=resid2), size = 3) + geom_hline(y=0, size = 2) + labs
p3 <- plot + geom_point(aes(x=seq_along(resid3) ,y=resid3), size = 3) + geom_hline(y=0, size = 2) + labs
grid.arrange(p1, p2, p3, ncol = 3, main = "Residual Plots")
```



Residual error has gradually decreased by adding the variables. Also, the predictors seem to be significant using maximum likelihood analysis.

```
anova(fit1, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am - 1
## Model 2: mpg ~ am + wt
## Model 3: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 68.734 5.071e-09 ***
## 3     28 180.29  1     98.03 15.224 0.0005464 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can reject the hypothesis that adding weight and horsepower is not required. Let us consider adding variables like cylinders and gears.

```
fit4 <- lm(mpg ~ am + wt + hp + cyl, data = cars)
fit5 <- lm(mpg ~ am + wt + hp + gear, data = cars)
anova(fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + wt + hp
## Model 2: mpg ~ am + wt + hp + cyl
```

```
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 180.29
## 2     27 170.00  1    10.293 1.6348 0.2119
```

```r
anova(fit3, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + wt + hp
## Model 2: mpg ~ am + wt + hp + gear
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 180.29
## 2     27 179.34  1   0.95072 0.1431 0.7081
```

We can reject the hypothesis that cylinders and gears contribute to outcome. Hence our final model does not include these. **We will stick with model 3**.

Conclusion with uncertainity quantified

```r
coeff <- summary(fit3)$coefficients
coeff[2,1] + c(-1,1) * coeff[2,2] * qt(0.975, fit3$df)
```

```
## [1] -0.7357587  4.9031790
```

```r
coeff[2,1] + c(-1,1) * coeff[2,2] * qt(0.925, fit3$df)
```

```
## [1] 0.04641094 4.12100932
```

For given weight and horse power, manaul transmission car is expected to give 2.08 more miles per gallon compared to a auto transmission car. With 95% confidence we can state that for the population of cars a manual transmission car gives 0.736 *less* miles per gallon to 4.9 *more* miles per gallon compared to automatic transmission car.

Executive summary (a.k.a no technical terms)

From the data, among cars with **same weight and horsepower**, we can answer the questions as follows

1. We can state with **85%** certainity that manual transmission cars have better mileage. However, by generally accepted statistic principles, we require a 95% certainity and hence can not make or publish this claim.
2. Manual transmission cars can give 0.736 *less* miles per gallon to 4.9 *more* miles per gallon compared to auto transmission cars