Motor Trend Cars Regression Model

Maurício Collaça

Nov 22, 2017

Executive Summary

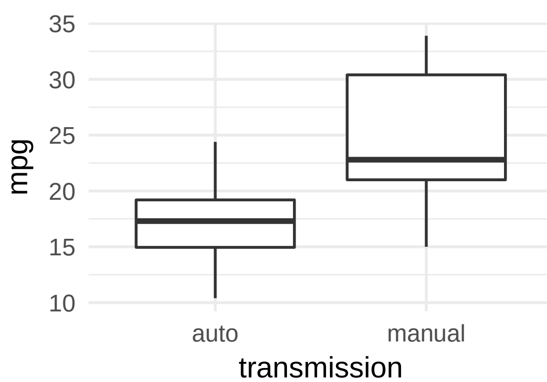
Motor Trend magazine is about the automobile industry and they are interested whether an automatic or manual transmission is better for MPG and quantify the MPG difference between automatic and manual transmissions. The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

It's concluded that manual transmission is better for gas mileage at a difference of 7.24 miles per galon.

Exploratory Data Analysis

The following boxplot suggests that manual transmission is better for miles per gallon and implies in a much higher variance.

```
ggplot(mtcars, aes(x=factor(am, labels = c("auto","manual")), y=mpg)) +
    geom_boxplot() + theme_minimal() + xlab("transmission")
```



Regression model selection

The strategy is to perform nested likelihood ratio tests starting from a base model `mpg ~ am` and adding regressors in the descending order of their correlation with `am`.

```
corMatrix <- cor(mtcars)
round(corMatrix[9,-c(1,9)][order(abs(corMatrix[9,-c(1,9)]),decreasing = TRUE)], 2)
```

```
 gear  drat    wt  disp   cyl    hp  qsec    vs  carb
 0.79  0.71 -0.69 -0.59 -0.52 -0.24 -0.23  0.17  0.06
```

The following model ANOVA suggests the first 4 models are adequate as they return p-values less than 0.05.

```
mtcars$cyl <- factor(mtcars$cyl); mtcars$vs <- factor(mtcars$vs)
mtcars$am <- factor(mtcars$am, labels = c("automatic", "manual"))
mtcars$gear <- factor(mtcars$gear); mtcars$carb <- factor(mtcars$carb)
fit1 <- lm(mpg ~ am, mtcars)
fit2 <- lm(mpg ~ am + gear, mtcars)
fit3 <- lm(mpg ~ am + gear + drat, mtcars)
fit4 <- lm(mpg ~ am + gear + drat + wt, mtcars)
fit5 <- lm(mpg ~ am + gear + drat + wt + disp, mtcars)
fit6 <- lm(mpg ~ am + gear + drat + wt + disp + cyl, mtcars)
fit7 <- lm(mpg ~ am + gear + drat + wt + disp + cyl + hp, mtcars)
fit8 <- lm(mpg ~ am + gear + drat + wt + disp + cyl + hp + qsec, mtcars)
fit9 <- lm(mpg ~ am + gear + drat + wt + disp + cyl + hp + qsec + vs, mtcars)
```

```r
fit10 <- lm(mpg ~ am + gear + drat + wt + disp + cyl + hp + qsec + vs + carb, mtcars)
as.data.frame(anova(fit1,fit2,fit3,fit4,fit5,fit6,fit7,fit8,fit9,fit10))
```

```
   Res.Df      RSS Df   Sum of Sq          F       Pr(>F)
1      30 720.8966 NA          NA         NA           NA
2      28 570.0023  2 150.894266  9.3993512 2.259094e-03
3      27 525.0707  1  44.931662  5.5976742 3.187411e-02
4      26 237.3032  1 287.767496 35.8506366 2.488053e-05
5      25 229.9164  1   7.386749  0.9202556 3.526165e-01
6      23 173.0313  2  56.885133  3.5434304 5.491354e-02
7      22 148.4018  1  24.629458  3.0683859 1.002422e-01
8      21 137.6778  1  10.723993  1.3360160 2.658171e-01
9      20 134.0015  1   3.676314  0.4580024 5.088617e-01
10     15 120.4027  5  13.598857  0.3388344 8.814442e-01
```

Inspecting coefficients of the first 4 models, one can see the first model results in the highest significant slope coefficient, as their p-values are the lowest ones and bellow 0.05, so the other models with additional regressors don't help to explain the data.

```r
summary(fit1)$coef; summary(fit2)$coef; summary(fit3)$coef; summary(fit4)$coef
```

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

               Estimate Std. Error     t value     Pr(>|t|)
(Intercept) 16.10666667   1.164967 13.82585505 4.917944e-14
ammanual     5.22500000   2.762962  1.89108635 6.900082e-02
gear4        4.94333333   2.538987  1.94697072 6.163258e-02
gear5        0.04833333   3.614215  0.01337312 9.894249e-01

             Estimate Std. Error    t value  Pr(>|t|)
(Intercept)  2.911933   8.754990  0.3326027 0.7420029
ammanual     4.082500   2.803142  1.4564016 0.1568131
gear4        1.869289   3.201284  0.5839184 0.5641222
gear5       -2.108552   3.806843 -0.5538848 0.5842162
drat         4.211981   2.771004  1.5200198 0.1401311

              Estimate Std. Error     t value     Pr(>|t|)
(Intercept) 33.2502250  8.0725686  4.11891516 3.426437e-04
ammanual     0.1247176  2.0456303  0.06096782 9.518511e-01
gear4        1.7520088  2.1932193  0.79882975 4.316263e-01
gear5       -1.3024026  2.6119213 -0.49863777 6.222270e-01
drat         0.5017393  2.0100561  0.24961457 8.048470e-01
wt          -4.8079279  0.8562531 -5.61507767 6.682894e-06
```

The residual plots in Appendix show the best model that explains the relatioship between the `mpg` and `am` is first because of is horizontal line in the Residuals vs Fitted plot and the diagonal fit in the Normal Q-Q plot.

The confidence intervals for the intercept and slope of the model 1 are:

```r
confint(fit1)
```

```
               2.5 %   97.5 %
(Intercept) 14.85062 19.44411
ammanual     3.64151 10.84837
```

Conclusions

Interpreting the model 1 coefficients one can see that the manual transmission is better for the gas milleage by increasing 7.24 miles per gallon. The intercept coefficient tells that the average miles per galon for automatic transmission is 17.15 with high significance as its p-value $1.1339835 \times 10^{-15}$ is less than 0.05. The slope coefficient `ammanual` tells that the increase in the average miles per galon for manual transmission is 7.24 with high significance as its p-value 2.8502074×10^{-4} is less than 0.05.

Appendix

Residual plots for models 1, 2, 3 and 4

```r
par(mfrow=c(4,4))
plot(fit1, which=c(1,2,3,5))
plot(fit2, which=c(1,2,3,5))
plot(fit3, which=c(1,2,3,5))
plot(fit4, which=c(1,2,3,5))
```