Analysis of MPG for Different Types of Transmission

Paul Ringsted, 31st January 2019 - Course 7 (Regression Models)

Synopsis

In this report we analyze whether the choice of automatic or manual transmission makes a difference to the Miles Per Gallon (MPG), using data from the 1974 Motor Trend US magazine for 32 different cars.

We conclude that cars with manual transmissions (which tend to have smaller engines and weigh less), do have a higher mean MPG than cars with automatic transmissions (which tend to have larger engines and weigh more). The difference in mean MPG is 7.2 +/- 4.0. There is not enough data to support a conclusion that the choice of transmission itself causes an MPG efficiency vs. other highly correlated factors such as engine size and vehicle weight. However, we are able to fit a regression model based on the interaction of weight and transmission to conclude that for each 1000lb decrease in vehicle weight, MPG improves by 9.1 for manual transmissions, vs. 3.8 for automatic transmissions.

Is an Automatic or Manual Transmission Better for MPG?

This analysis is on the R mtcars dataset. This 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). We first gather some basic statistics to understand the MPG data for different transmissions:

Transmission	MPG Mean	MPG Min	MPG Max	MPG SD	Cars	4 Cyl	6 Cyl	8 Cyl
Automatic	17.15	10.4	24.4	3.834	19	3	4	12
Manual	24.39	15.0	33.9	6.167	13	8	3	2

Plots of the data are shown in the Appendix:

- Figure 1 - Boxplot of MPG by transmission type
- Table 3 - Correlation matrix for MPG data
- Figure 3 - Plot of MPG by Weight, by Cylinder and Transmission (with regression model)

The highest correlations between MPG and other variables are with displacement, cylinders and weight. There is a high degree of correlation between the variables e.g. displacement is a function of cylinders therefore highly correlated; heavier cars have bigger engines with higher displacements, etc.

As shown in the data table and illustrated in Figure 3, the transmission type is not well randomized across cars with different weights and engine types. We can see a clear grouping of “small” cars (4 cylinders, weight <= 3.165, 10-20 MPG) having predominantly manual transmissions; “big” cars (8 cylinders, weight >= 3.165, 20-35 MPG) having predominantly automatic transmissions; and a cluster of “midsize” cars (6 cylinders, weight 2.5-3.5, 17-22 MPG) roughly evenly split between manual and automatic transmissions.

Based on Figure 1, looking purely at transmission type and ignoring all other car features, the 25-75% quantiles for the populations are distinct with a higher MPG evident in cars with manual transmissions, however there is a large degree of overlap.

To statistically confirm whether transmission type significantly affects mean MPG, we will perform a two-sided 95% t-test between the different transmission types. Our null hypothesis is that transmission type has no impact on mean MPG i.e.

Null hypothesis $H_0 : \mu_{auto} - \mu_{man} = 0$ vs. alternative $H_a : \mu_{auto} - \mu_{man} \neq 0$

Table 1: Results of T-Test in Mean MPG Between Transmissions

	T-Test	T-Statistic	DoF	Conf Int Low	Conf Int High	P-Value %
t	Auto vs. Manual, All Cylinders	-3.767	18	-11.28	-3.21	0.00137

Across all type of car, the resulting Confidence Interval does not bound 0, and has a significant P-Value (0.0014) < 0.05.

Conclusion: Cars with manual transmissions, have a higher mean MPG than cars with automatic transmissions. The difference in mean MPG is 7.24 ± 4.04.

Modeling MPG

See appendix for exploratory model fitting output. Fitting a basic model with all variables, anova() highlights the variables cyl, disp and wt as being significant. Comparing these for fit using anova() shows that adding disp adds no value, and there are two viable options for the model with roughly equivalent RSS which were considered:

- mpg~cyl+wt $\implies$ 3 regression lines with same slope but different intercepts by cylinder (as a factor)
- mpg~wt*am $\implies$ 2 regression lines with different slopes by transmission type (as a factor)

Since we want to gain insight on the effect of transmission type on MPG, the second model was selected, ie our model is as follows:

$MPG_i = \beta_0 + \beta_1 WT_i + \beta_2 AM_i + \beta_3 WT_i AM_i + e_i$, where:

For car i: MPG_i= MPG; WT_i = Weight; AM_i = Transmission (0/1); e_i = Error of the model

$AM_i = 0 \implies MPG_i = \beta_0 + \beta_1 WT_i + e_i$ (Automatic)

$AM_i = 1 \implies MPG_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times WT_i + e_i$ (Manual)

Table 2: Coefficients and CI of MPG Model (wt*am)

	Estimate	Std. Error	t value	Pr(>|t|)	2.5 %	97.5 %
(Intercept)	31.416055	3.0201093	10.402291	0.0000000	25.229642	37.602469
wt	-3.785907	0.7856478	-4.818836	0.0000455	-5.395234	-2.176581
amManual	14.878422	4.2640422	3.489276	0.0016210	6.143928	23.612917
wt:amManual	-5.298361	1.4446993	-3.667449	0.0010171	-8.257693	-2.339028

This model yields 2 regression lines; for automatic transmissions the slope is β_1 = -3.79, and for manual transmissions the slope is $\beta_1 + \beta_3$ = -9.08. The slopes represent the rate by which MPG changes by weight of vehicle.

Figure 3 reflects the regression lines. Figure 4 shows the fit diagnostics including residuals vs. fit which looks reasonable. Note the two outliers (Toyota Corolla and Fiat 128) which have the highest MPGs; their removal would further improve model fit but given small size of the dataset no exclusions were applied. The model has RSS = 188, $R^2 = 0.833$ and all estimates lie outside 2 s.d. of zero with low p-values, so this appears to be a reasonable model fit.

Conclusion: For each 1000lb change in vehicle weight, MPG changes by -9.08 for manual transmissions, and -3.79 for automatic transmissions.

Appendix - Figures and Tables

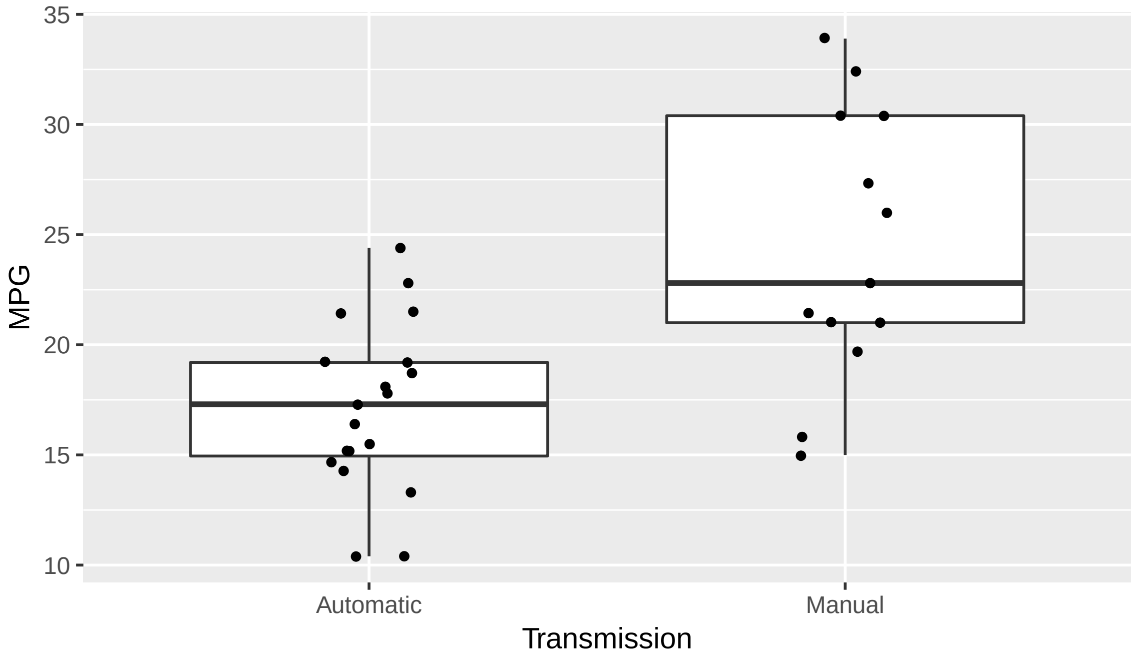

Figure 1: MPG by Transmission

Table 3: Correlation Matrix for MPG Data

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	100.0	-85.2	-84.8	-77.6	68.1	-86.8	41.9	66.4	60.0	48.0	-55.1
cyl	-85.2	100.0	90.2	83.2	-70.0	78.2	-59.1	-81.1	-52.3	-49.3	52.7
disp	-84.8	90.2	100.0	79.1	-71.0	88.8	-43.4	-71.0	-59.1	-55.6	39.5
hp	-77.6	83.2	79.1	100.0	-44.9	65.9	-70.8	-72.3	-24.3	-12.6	75.0
drat	68.1	-70.0	-71.0	-44.9	100.0	-71.2	9.1	44.0	71.3	70.0	-9.1
wt	-86.8	78.2	88.8	65.9	-71.2	100.0	-17.5	-55.5	-69.2	-58.3	42.8
qsec	41.9	-59.1	-43.4	-70.8	9.1	-17.5	100.0	74.5	-23.0	-21.3	-65.6
vs	66.4	-81.1	-71.0	-72.3	44.0	-55.5	74.5	100.0	16.8	20.6	-57.0
am	60.0	-52.3	-59.1	-24.3	71.3	-69.2	-23.0	16.8	100.0	79.4	5.8
gear	48.0	-49.3	-55.6	-12.6	70.0	-58.3	-21.3	20.6	79.4	100.0	27.4
carb	-55.1	52.7	39.5	75.0	-9.1	42.8	-65.6	-57.0	5.8	27.4	100.0

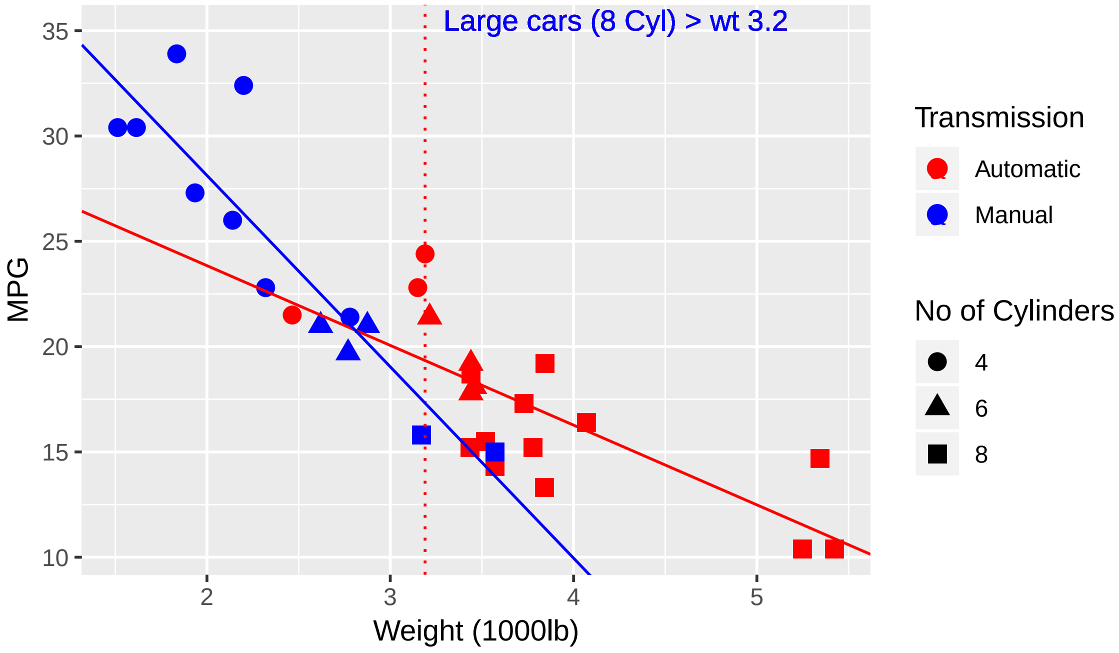

Figure 2: MPG by Weight with Model Fit

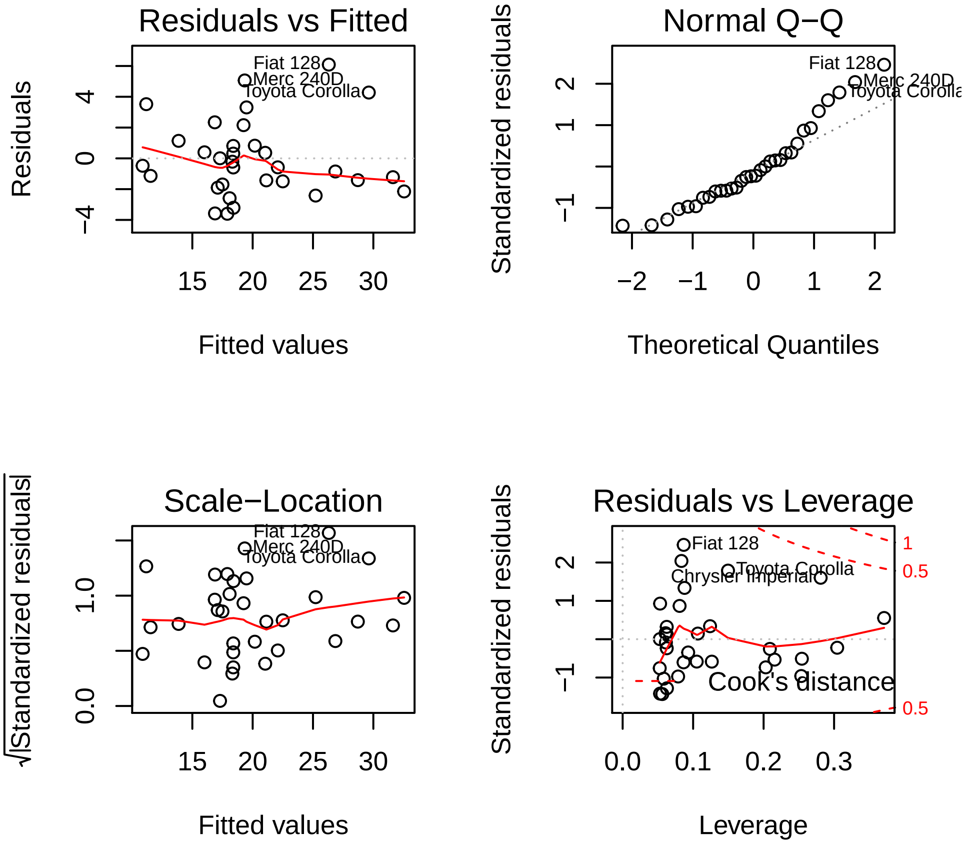

Figure 3: Model Fit Diagnostics

Anova for broad model with all variables lm(mpg~.,mt):

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## cyl        2 824.78  412.39 51.3766 1.943e-07 ***
## disp       1  57.64   57.64  7.1813   0.01714 *
## hp         1  18.50   18.50  2.3050   0.14975
## drat       1  11.91   11.91  1.4843   0.24191
## wt         1  55.79   55.79  6.9500   0.01870 *
## qsec       1   1.52    1.52  0.1899   0.66918
## vs         1   0.30    0.30  0.0376   0.84878
## am         1  16.57   16.57  2.0639   0.17135
## gear       2   5.02    2.51  0.3128   0.73606
## carb       5  13.60    2.72  0.3388   0.88144
## Residuals 15 120.40    8.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova fits for cyl, wt, disp and am combinations:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ cyl
## Model 2: mpg ~ cyl + wt
## Model 3: mpg ~ cyl + wt + disp
## Model 4: mpg ~ cyl + wt + disp + am
## Model 5: mpg ~ cyl + wt + am
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     29 301.26
## 2     28 183.06  1   118.204 16.8060 0.0003605 ***
## 3     27 182.95  1     0.110  0.0156 0.9015995
## 4     26 182.87  1     0.080  0.0113 0.9160547
## 5     27 182.97 -1    -0.099  0.0141 0.9064703
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Anova fits for wt, am combinations:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + am
## Model 3: mpg ~ wt * am
##   Res.Df    RSS Df Sum of Sq       F   Pr(>F)
## 1     30 278.32
## 2     29 278.32  1     0.002  0.0003 0.985556
## 3     28 188.01  1    90.312 13.4502 0.001017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Code Appendix - R Code

```r
library(knitr)
opts_chunk$set(fig.width=6, fig.height=3.5, fig.pos = "H", echo=FALSE, eval=TRUE)
library(ggplot2)
library(kableExtra)
library(datasets)

#--------------------------------------------------------------------------------------
# Load mtcars data and copy to 'mt' dataframe for ease of reference
data(mtcars)
mt<-mtcars

# Clean up flags to make them factor variables with proper names
for (i in c(2,8:11)) { mt[,i]<-as.factor(mt[,i])}
levels(mt$am)<-c("Automatic","Manual")

# Build and render a table of basic stats for the dataset
mtstats<-cbind(aggregate(mt$mpg,by=list(mt$am),FUN=mean))
mtstats<-cbind(mtstats,aggregate(mt$mpg,by=list(mt$am),FUN=min)$x)
mtstats<-cbind(mtstats,aggregate(mt$mpg,by=list(mt$am),FUN=max)$x)
mtstats<-cbind(mtstats,aggregate(mt$mpg,by=list(mt$am),FUN=sd)$x)

cyltab<-table(mt$am,mt$cyl)
mtstats<-cbind(mtstats,aggregate(mt$mpg,by=list(mt$am),FUN=length)$x)
mtstats<-cbind(mtstats,rbind(cyltab[1,],cyltab[2,]))

mtstats %>% kable(col.names=c("Transmission","MPG Mean","MPG Min","MPG Max","MPG SD",
                "Cars","4 Cyl","6 Cyl","8 Cyl"),booktabs=T,
                align=c("c",rep("r",8)),digits=c(2,2,2,2,3,2,2,2,2)) %>%
        kable_styling(latex_options = "hold_position")

#--------------------------------------------------------------------------------------
# Perform t-test for mean MPG by transmission type

test1<-t.test(mpg~am,paired=FALSE,var.equal=FALSE,data=mt)
c_lo<-test1$conf.int[1]
c_hi<-test1$conf.int[2]
pval<-test1$p.value
tval<-test1$statistic
dval<-test1$parameter
testtype<-"Auto vs. Manual, All Cylinders"

#Build a dataframe with the test statistics (p-values in % for display) and render it
results<-data.frame(testtype,tval,dval,c_lo,c_hi,pval,stringsAsFactors = FALSE)
results %>% kable(col.names=c("T-Test","T-Statistic","DoF","Conf Int Low",
                              "Conf Int High","P-Value %"),
                  align=c("l",rep("r",5)),digits=c(2,3,0,2,2,5),booktabs=T,
                  caption="Results of T-Test in Mean MPG Between Transmissions") %>%
                kable_styling(latex_options = "hold_position")

ci<-(c_hi-c_lo)/2
ci_mid<-abs(c_lo+ci)
```

```
#--------------------------------------------------------------------------------------
# Fit model to cyl+wt and get coefficients to use in report and plots
fitmt<-lm(mpg~wt*am,mt)
beta0<-coef(fitmt)[1]
beta1<-coef(fitmt)[2]
beta2<-coef(fitmt)[3]
beta3<-coef(fitmt)[4]
itc_auto<-beta0
itc_man<-beta0+beta2
slope_auto<-beta1
slope_man<-beta1+beta3
rss<-sum(fitmt$residuals^2)
r2<-summary(fitmt)$r.squared
call<-summary(fitmt)$call
coef<-summary(fitmt)$coefficients
ci<-confint(fitmt)
coef2<-cbind(coef,as.data.frame(ci))
coef2 %>% kable(caption="Coefficients and CI of MPG Model (wt*am)", booktabs=T) %>%
                kable_styling(latex_options = "hold_position")

#--------------------------------------------------------------------------------------
# Fig1: Box plot of MPG by transmission type
g <- ggplot(data=mt,aes(x=am,y=mpg))
g <- g + geom_boxplot(aes(group=am))
g <- g + geom_jitter(shape=16, position=position_jitter(0.1))
g <- g + labs(x="Transmission",y="MPG")
print(g)
#--------------------------------------------------------------------------------------
# Table3: Correlation data (using original dataframe)
cor_res<-round(100*cor(mtcars),1)
cor_res %>% kable(booktabs=T,
                caption="Correlation Matrix for MPG Data") %>%
                kable_styling(latex_options = "hold_position")
#--------------------------------------------------------------------------------------
# Fig2: Plot of MPG by weight with model fit
wtcut<-max(subset(mt,cyl == 4)$wt)
g <- ggplot(data=mt,aes(x=wt,y=mpg,col=am,pch=cyl))
g <- g + geom_point(size=3)
g <- g + geom_vline(xintercept=wtcut,col="red",linetype="dotted")
g <- g + geom_abline(intercept=itc_auto,slope=slope_auto,col="red")
g <- g + geom_abline(intercept=itc_man,slope=slope_man,col="blue")
g <- g + geom_text(aes(wtcut+0.1, 35, vjust=0, hjust=0,
                        label = paste0("Large cars (8 Cyl) > wt ",round(wtcut,1))))
g <- g + labs(x="Weight (1000lb)",y="MPG",pch="No of Cylinders",color="Transmission")
g <- g + scale_colour_manual(values = c("red","blue"))
print(g)

#--------------------------------------------------------------------------------------
# Fig3: Plot of model fit diagnostics
par(mfrow=c(2,2))
plot(fitmt)

#--------------------------------------------------------------------------------------
```

```r
# Model selection - anova() results for all variables
fitall<-lm(mpg~.,mt)
anova(fitall)

#--------------------------------------------------------------------------------------
# Model selection - anova() results for subset of significant variables
fit1<-lm(mpg~cyl,mt)
fit2<-lm(mpg~cyl+wt,mt)
fit3<-lm(mpg~cyl+wt+disp,mt)
fit4<-lm(mpg~cyl+wt+disp+am,mt)
fit5<-lm(mpg~cyl+wt+am,mt)
anova(fit1,fit2,fit3,fit4,fit5)

#--------------------------------------------------------------------------------------
# Model selection - anova() results for subset of significant variables
fit6<-lm(mpg~wt,mt)
fit7<-lm(mpg~wt+am,mt)
fit8<-lm(mpg~wt*am,mt)
anova(fit6,fit7,fit8)
```