Regression Models Course Project

Daria Karpova

4/26/2021

About this project

This analysis is performed for the Regression Models course by John Hopkins University on Coursera. Th data used for analysis is the built-in 'mtcars' R dataset. The data contains 32 observations on 11 features.

- **mpg**: Miles per gallon
- **cyl**: Number of cylinders
- **disp**: Displacement (cu.in.)
- **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

Instructions

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

1. "Is an automatic or manual transmission better for MPG"
2. "Quantify the MPG difference between automatic and manual transmissions"

Loading data

```
library(nlcor)
library(knitr)
library(datasets)
data(mtcars)
```

Exploring data

```
# Variable summary
kable(summary(mtcars[1:10]))
```

mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear
Min. :10.40	Min. :4.000	Min. : 71.1	Min. : 52.0	Min. :2.760	Min. :1.513	Min. :14.50	Min. :0.0000	Min. :0.0000	Min. :3.000
1st Qu.:15.43	1st Qu.:4.000	1st Qu.:120.8	1st Qu.: 96.5	1st Qu.:3.080	1st Qu.:2.581	1st Qu.:16.89	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:3.000
Median :19.20	Median :6.000	Median :196.3	Median :123.0	Median :3.695	Median :3.325	Median :17.71	Median :0.0000	Median :0.0000	Median :4.000
Mean :20.09	Mean :6.188	Mean :230.7	Mean :146.7	Mean :3.597	Mean :3.217	Mean :17.85	Mean :0.4375	Mean :0.4062	Mean :3.688
3rd Qu.:22.80	3rd Qu.:8.000	3rd Qu.:326.0	3rd Qu.:180.0	3rd Qu.:3.920	3rd Qu.:3.610	3rd Qu.:18.90	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:4.000
Max. :33.90	Max. :8.000	Max. :472.0	Max. :335.0	Max. :4.930	Max. :5.424	Max. :22.90	Max. :1.0000	Max. :1.0000	Max. :5.000

```
#Taking a look at the data
kable(head(mtcars))
```

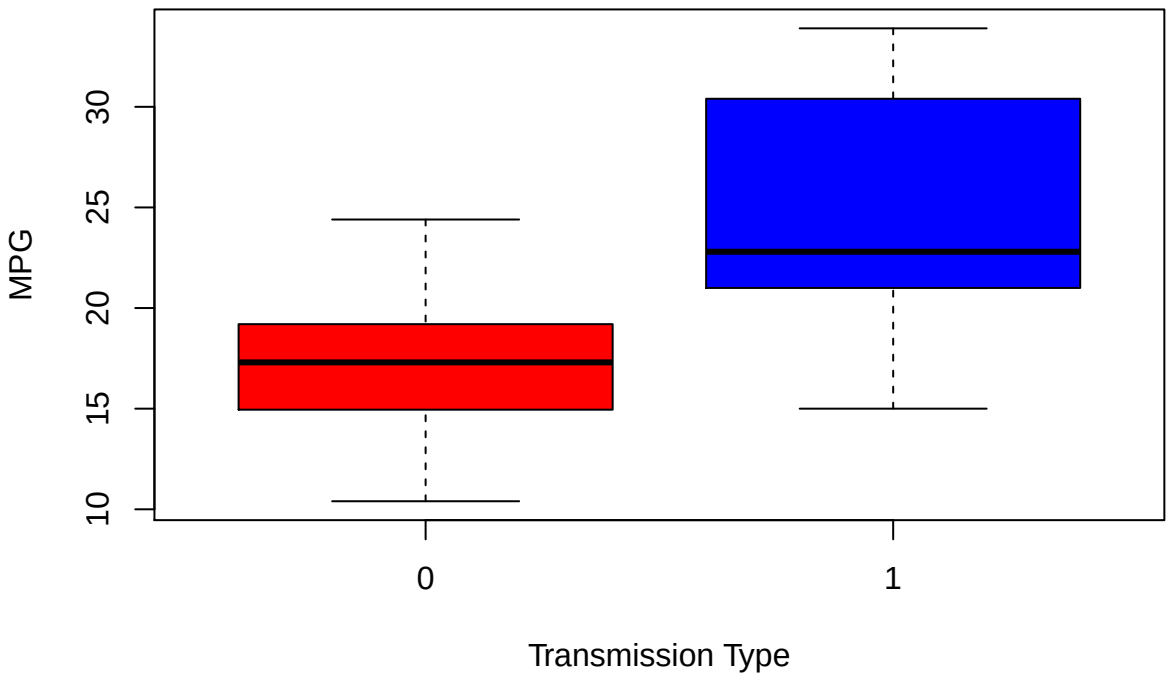
	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Visualization

Before applying any model let's see how the data looks on a graph.

```
# For easier representation
levels(mtcars$am) <- c("Automatic", "Manual")

plot(mpg ~ as.factor(am), data = mtcars, col = (c("red","blue")), ylab = "MPG", xlab = "Transmission Ty
```



It can be seen that Manual Transmission has noticeably better MPG. Seems like there is a dependency between them.

Testing for transmission type equality

Let's check if the Automatic and Manual transmission differ by using a t-test. Our hypothesis is that Automatic transmission MPG mean is equal to the Manual Transmission MPG mean.

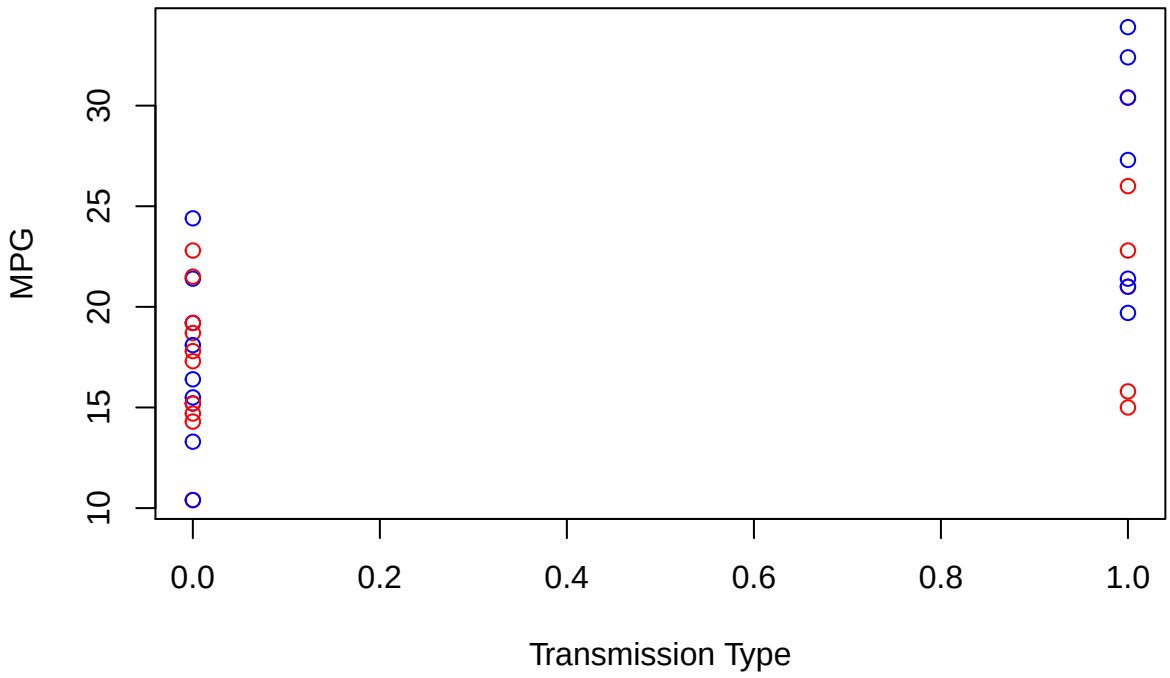
```
t.test(mpg ~ am, mtcars, alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.9993
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -10.57662       Inf
## sample estimates:
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

The hypothesis can be rejected with 99.3% confidence, therefore, it is clear that Manual transmission is better for MPG.

Simple model visualization

```
plot(mpg ~ am, data = mtcars, col = (c("red","blue")), ylab = "MPG", xlab = "Transmission Type")
```



From the above graph it is obvious that a linear function of the transmission type alone won't be a good choice. MPG can not be predicted by a single binary value.

Univariate regression

To confirm that let's see what would we get from a univariate model depending only on the transmission type.

```
linear_univariate <- lm(mpg ~ am, data = mtcars)
summary(linear_univariate)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## 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:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

R-squared is only 0.3385, the model underfits the data. Hence, we have to look for other related variables.

Finding correlations

```
mtcars$am <- as.numeric(mtcars$am)
cor_matrix <- cor(mtcars)

#Printing all correlation coeffitions between MPG and other mtcars variables
kable(cor_matrix[2:11,1], col.names='MPG')
```

	MPG
cyl	-0.8521620
disp	-0.8475514
hp	-0.7761684
drat	0.6811719
wt	-0.8676594
qsec	0.4186840
vs	0.6640389
am	0.5998324
gear	0.4802848
carb	-0.5509251

The correlation table suggests us to use 'cyl', 'disp' and 'wt' variables as they have strong correlation. At the same time Transmission type does not highly correlate with MPG.

Multivariate regression

Let's try to fit the data with a simple first-order regression.

```
linear_multivariate <- lm(mpg ~ cyl + disp + wt + am, data = mtcars)
summary(linear_multivariate)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.318 -1.362 -0.479  1.354  6.059
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 40.898313   3.601540  11.356 8.68e-12 ***
## cyl         -1.784173   0.618192  -2.886  0.00758 **
## disp         0.007404   0.012081   0.613  0.54509
## wt          -3.583425   1.186504  -3.020  0.00547 **
## am           0.129066   1.321512   0.098  0.92292
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.642 on 27 degrees of freedom
## Multiple R-squared:  0.8327, Adjusted R-squared:  0.8079
## F-statistic: 33.59 on 4 and 27 DF,  p-value: 4.038e-10
```

Adjusted R-squared is 80.79%. What if we also use 'hp' and 'vs'?

```
linear_multivariate <- lm(mpg ~ cyl + disp + wt + am + hp + vs, data = mtcars)
summary(linear_multivariate)
```

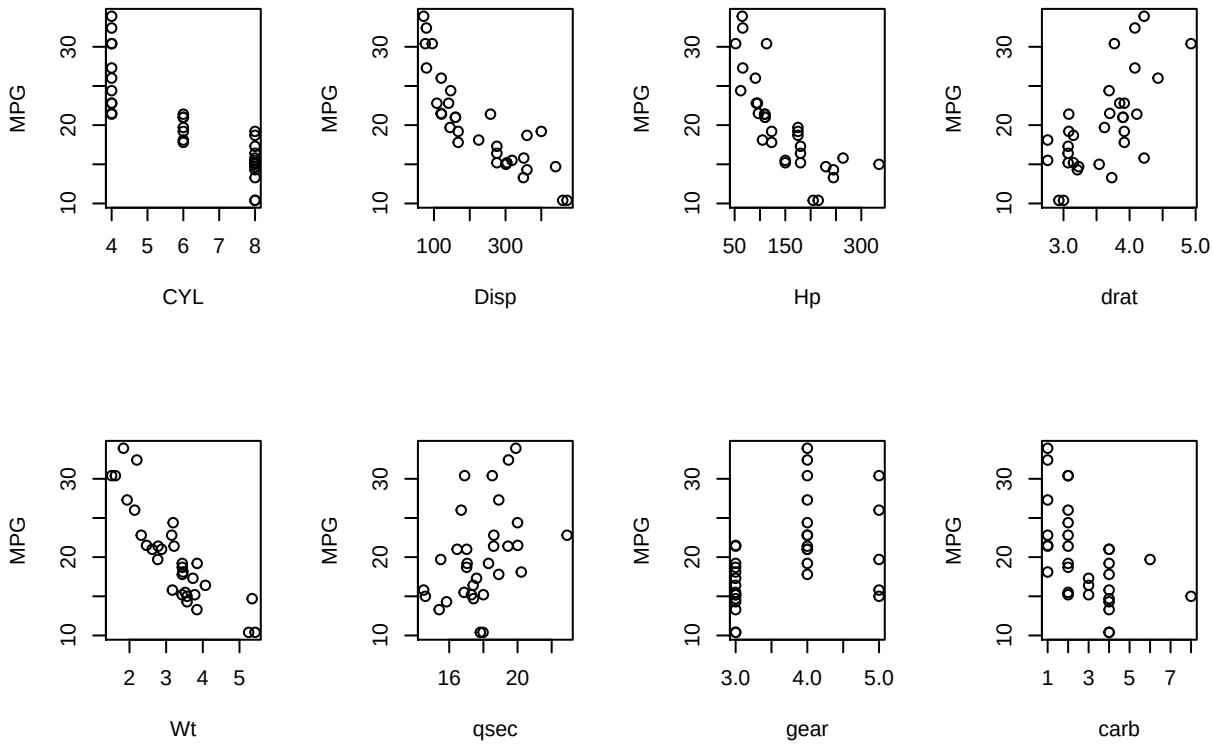
```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am + hp + vs, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8319 -1.7327 -0.4034  1.3154  5.3430
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.96611    5.68062   6.155 1.95e-06 ***
## cyl         -0.73198    0.84488  -0.866  0.39452
## disp         0.01311    0.01186   1.105  0.27964
## wt          -3.27739    1.14376  -2.865  0.00832 **
## am           2.14088    1.64807   1.299  0.20579
## hp          -0.02926    0.01415  -2.068  0.04911 *
## vs           1.36178    1.81343   0.751  0.45970
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.526 on 25 degrees of freedom
## Multiple R-squared:  0.8583, Adjusted R-squared:  0.8243
## F-statistic: 25.25 on 6 and 25 DF,  p-value: 1.817e-09
```

Now it's 82.43%, so generally those variables do not influence the result much when using a simple multivariate model.

Finding non-linear dependencies

Let's now check the scatterplot for any non-linear dependencies.

```
par(mfrow=c(2,4))
plot(mpg ~ cyl, data = mtcars, ylab = "MPG", xlab = "CYL")
plot(mpg ~ disp, data = mtcars, ylab = "MPG", xlab = "Disp")
plot(mpg ~ hp, data = mtcars, ylab = "MPG", xlab = "Hp")
plot(mpg ~ drat, data = mtcars, ylab = "MPG", xlab = "drat")
plot(mpg ~ wt, data = mtcars, ylab = "MPG", xlab = "Wt")
plot(mpg ~ qsec, data = mtcars, ylab = "MPG", xlab = "qsec")
plot(mpg ~ gear, data = mtcars, ylab = "MPG", xlab = "gear")
plot(mpg ~ carb, data = mtcars, ylab = "MPG", xlab = "carb")
```



Seems like 'hp' might have a quadratic dependency. As well as 'disp'.

```
mtcars$hp_squared = (mtcars$hp)^2
mtcars$disp_squared = (mtcars$disp)^2
```

Now let's fit it into the model.

```
linear_multivariate <- lm(mpg ~ cyl + disp + disp_squared + wt + am + hp + hp_squared, data = mtcars)
summary(linear_multivariate)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + disp_squared + wt + am + hp +
##     hp_squared, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2489 -1.4544 -0.1489  1.5839  3.5713
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.532e+01  4.114e+00  11.014 7.21e-11 ***
## cyl           4.957e-01  8.296e-01   0.598  0.55572
## disp         -6.419e-02  3.608e-02  -1.779  0.08792 .
## disp_squared  1.130e-04  5.091e-05   2.219  0.03617 *
## wt           -3.670e+00  1.049e+00  -3.500  0.00184 **
## am           -6.283e-01  1.560e+00  -0.403  0.69067
## hp           -9.383e-02  4.282e-02  -2.191  0.03839 *
## hp_squared    1.782e-04  9.747e-05   1.828  0.07998 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.247 on 24 degrees of freedom
## Multiple R-squared:  0.8924, Adjusted R-squared:  0.861
## F-statistic: 28.44 on 7 and 24 DF,  p-value: 3.898e-10
```

The final result is 86.1% for Adjusted R-squared.

Transmission type influence

'am's p-value is the highest here that indicates it doesn't influence the result much. Let's check for correlation with other variables in the model. Maybe the information 'am' was providing is now obtained from a different variable.

```
cor_matrix <- cor(mtcars)

#Printing all correlation coeffitions between MPG and other mtcars variables
kable(cor_matrix[1:11, 9], col.names='AM')
```

	AM
mpg	0.5998324
cyl	-0.5226070
disp	-0.5912270
hp	-0.2432043
drat	0.7127111
wt	-0.6924953
qsec	-0.2298609
vs	0.1683451
am	1.0000000
gear	0.7940588
carb	0.0575344

'wt', 'cyl' and 'disp' column have the highest correlation with am, we'll try to remove them one by one.

```r
linear_multivariate <- lm(mpg ~ cyl + disp + disp_squared + am + hp + hp_squared, data = mtcars)
summary(linear_multivariate)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + disp_squared + am + hp + hp_squared,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6771 -1.8389 -0.4082  1.3973  5.5280
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.612e+01  3.813e+00   9.474 9.43e-10 ***
## cyl           4.649e-01  9.988e-01   0.465   0.6457
## disp         -5.745e-02  4.338e-02  -1.324   0.1974
## disp_squared  7.226e-05  5.968e-05   1.211   0.2373
## am            1.989e+00  1.648e+00   1.207   0.2387
## hp           -1.145e-01  5.107e-02  -2.243   0.0340 *
## hp_squared    2.078e-04  1.169e-04   1.777   0.0877 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.705 on 25 degrees of freedom
## Multiple R-squared:  0.8375, Adjusted R-squared:  0.7985
## F-statistic: 21.47 on 6 and 25 DF,  p-value: 9.653e-09
```

```r
# Change in MPG when changing to manual
(summary(linear_multivariate)$coefficient['am', 'Estimate'])# * sd_mpg + mean_mpg
```

```
## [1] 1.989462
```

```r
linear_multivariate <- lm(mpg ~ disp + disp_squared + am + hp + hp_squared, data = mtcars)
summary(linear_multivariate)
```

```
##
## Call:
## lm(formula = mpg ~ disp + disp_squared + am + hp + hp_squared,
##     data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6728 -1.7636 -0.4025  1.3269  5.3220
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.610e+01  3.755e+00   9.616 4.76e-10 ***
## disp         -4.372e-02  3.133e-02  -1.395   0.1747
## disp_squared  5.528e-05  4.651e-05   1.188   0.2454
## am            2.165e+00  1.580e+00   1.371   0.1822
## hp           -1.058e-01  4.678e-02  -2.262   0.0323 *
```

```
## hp_squared    1.900e-04  1.088e-04   1.746   0.0926 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.664 on 26 degrees of freedom
## Multiple R-squared:  0.8361, Adjusted R-squared:  0.8046
## F-statistic: 26.53 on 5 and 26 DF,  p-value: 1.946e-09
```

```
# Change in MPG when changing to manual
(summary(linear_multivariate)$coefficient['am', 'Estimate'])# * sd_mpg + mean_mpg
```

```
## [1] 2.165272
```

```
linear_multivariate <- lm(mpg ~ disp_squared + am + hp + hp_squared, data = mtcars)
summary(linear_multivariate)
```

```
##
## Call:
## lm(formula = mpg ~ disp_squared + am + hp + hp_squared, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2440 -1.8735 -0.1678  1.1192  6.3011
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.340e+01  3.274e+00  10.204  9.2e-11 ***
## disp_squared -6.892e-06  1.357e-05  -0.508  0.61573
## am            3.498e+00  1.280e+00   2.732  0.01096 *
## hp           -1.398e-01  4.064e-02  -3.440  0.00191 **
## hp_squared    2.395e-04  1.046e-04   2.288  0.03016 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.711 on 27 degrees of freedom
## Multiple R-squared:  0.8238, Adjusted R-squared:  0.7977
## F-statistic: 31.56 on 4 and 27 DF,  p-value: 8.015e-10
```

```
# Change in MPG when changing to manual
(summary(linear_multivariate)$coefficient['am', 'Estimate'])# * sd_mpg + mean_mpg
```

```
## [1] 3.498002
```

Surely, R-squared decreased to 79.77% since we removed 'wt', 'cyl' and 'disp' but during each step we improved our confidence in transmission type having effect on the result. Finally, our confidence reached almost 99% so we'll use that model to quantify the difference. Since the indepent variable of interest is binary estimate corresponds to the change in MPG (when replacing 0 (automatic) by 1 (manual)). We can conclude respectively that manual transmission improved MPG by 3.498.

Conclusion

1) Manual transmission is better for MPG than automatic one.
2) Changing to manual improves MPG by 3.398.