Regression Modeling in R: Impact of Car Transmission Type (Automatic vs. Manual) on Mileage Based on the "Motor Trends" Cars Dataset

Stefan Schmager

April 3, 2016

Author's Remark

The author of this paper is aware of the page limitation, but kindly asks to excuse that violation. This project was way too much fun to be pressed into two pages! ;)

Executive Summary

This is the final project for the course Regression Models hosted by the Johns Hopkins University on Coursera as part of the Data Science Specialization. The project is based on the widely demonstrated R dataset *mtcars*, which is an extract from an article of *Motor Trend*, a magazine about the automobile industry. Looking at a data set of a collection of cars, the project aims to explore the relationship between a set of predictor variables (car features) and miles per gallon (MPG) as the outcome variable. One is particularly interested in the following two questions: **Is an automatic or manual transmission better for MPG? How can the MPG difference between automatic and manual transmissions be quantified?** After preparing the data and applying a few data manipulations, the two main variables (both predictor of interest and outcome) are visualized by a boxplot as part of an initial exploratory data analysis (EDA). Further, the EDA is translated into a first, linear regression model that's unadjusted for any other variable. Then, the basic multivariate regression model is constructed by including all other available (and potentially MPG predicting) variables before a nested regression modeling approach is applied. This approach always includes transmission type as a steady regressor, but gradually adjusts for other selected variables. Both from a statistical and logical, common-sense point of view certain variables are added to the model sequence and removed from the model after the series of regression models has been tested with the help of analysis of variance. The final model that's used to answer the research questions is summarized and troughly explained with the help of coeficient, other summary-measure, and diagnostic-plot interpretations. It turns out that neither automatic, nor manual transmissions are better for MPG. Rather, other features of a car can predict its MPG much better than solely the transmission type.

Data Preparation

```
library(datasets); library(ggplot2); library(dplyr)
```

These R packages are used in the whole analysis.

```
data(mtcars) #Load data set
```

The *Motor Trend* dataset (mtcars) is the basis of this analysis. According to its R documenation, the data was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption, miles per (U.S.)

gallon (_MPG___), and 10 additional aspects of automobile design and performance (such as automatic or manual transmission, abbr. **Trans**) for 32 automobiles (1973-74 models).

```
cars <- transmute(mtcars,
                  MPG    = as.numeric(mpg),
                  Trans  = factor(am,
                                  levels = c(0,      1),
                                  labels = c("Auto", "Manu")),
                  Weight = as.numeric(wt),       # Weight (1000 lbs)
                  Cyl    = as.numeric(cyl),      # Number of cylinders
                  Displ  = as.numeric(disp),     # Displacement (cu.in.)
                  HP     = as.numeric(hp),       # Gross horsepower
                  RAR    = as.numeric(drat),     # Rear axle ratio
                  QMTime = as.numeric(qsec),     # 1/4 mile time
                  VS     = factor(vs),           # ??? (documentation doesn't explain)
                  Gear   = as.numeric(gear),     # Number of forward gears
                  Carb   = as.numeric(carb)) %>% # Number of carburetors
      arrange(desc(MPG))
row.names(cars) <- row.names(mtcars)             # Name row names
head(cars, 5)                                    # Top 5 cars re: MPG
```

```
##                    MPG Trans Weight Cyl Displ  HP  RAR QMTime VS Gear Carb
## Mazda RX4         33.9  Manu  1.835   4  71.1  65 4.22  19.90  1    4    1
## Mazda RX4 Wag     32.4  Manu  2.200   4  78.7  66 4.08  19.47  1    4    1
## Datsun 710        30.4  Manu  1.615   4  75.7  52 4.93  18.52  1    4    2
## Hornet 4 Drive    30.4  Manu  1.513   4  95.1 113 3.77  16.90  1    5    2
## Hornet Sportabout 27.3  Manu  1.935   4  79.0  66 4.08  18.90  1    4    1
```

```
tail(cars, 5)                                    # Bottom 5 cars re: MPG
```

```
##                 MPG Trans Weight Cyl Displ  HP  RAR QMTime VS Gear Carb
## Lotus Europa   14.7  Auto  5.345   8   440 230 3.23  17.42  0    3    4
## Ford Pantera L 14.3  Auto  3.570   8   360 245 3.21  15.84  0    3    4
## Ferrari Dino   13.3  Auto  3.840   8   350 245 3.73  15.41  0    3    4
## Maserati Bora  10.4  Auto  5.250   8   472 205 2.93  17.98  0    3    4
## Volvo 142E     10.4  Auto  5.424   8   460 215 3.00  17.82  0    3    4
```

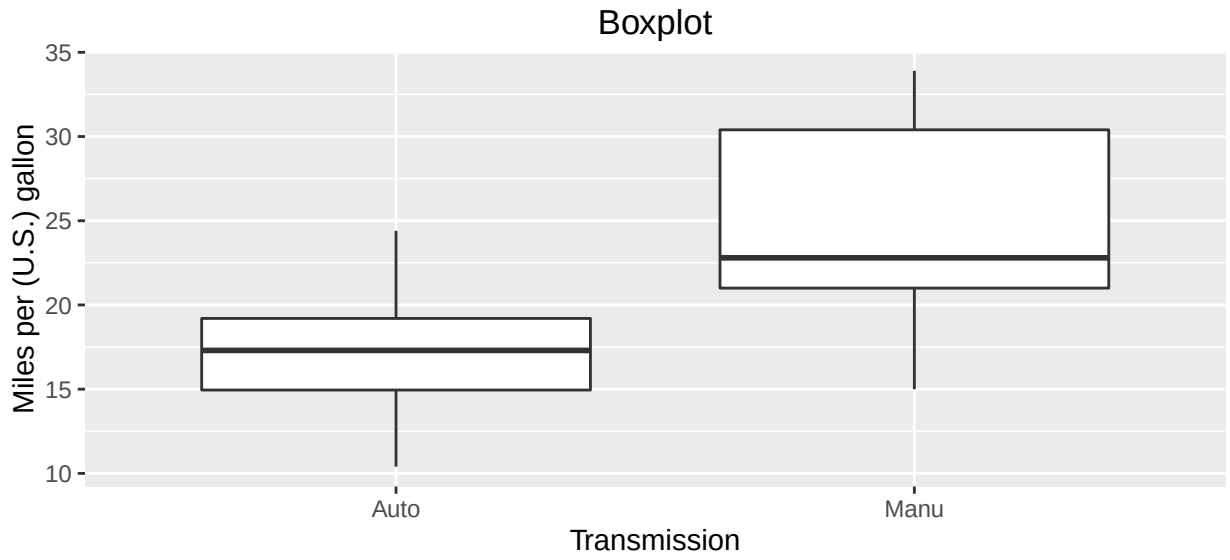
```
table(cars$Trans)                                # Frequency table re: Trans
```

```
##
## Auto Manu
##   19   13
```

The original dataset is transformed to an information-identical dataset, however, with more intuitive variable names and data types. **MPG** remains a continuous variable, whereas **Trans** (formerly *am* in original dataset) is transformed to a nominal factor variale with 0 being automatic (**auto**) and 1 being manual (**manu**). The 32 cars are listed in descending order by their mileage and indicated by their transmission type: 19 cars are automatic, 13 are manual. The top and bottom five most, respectively least fuel-efficient cars are shown.

Exploratory Data Analysis

```
ggplot(cars, aes(x = Trans, y = MPG)) + geom_boxplot() +
      labs(title = "Boxplot", x = "Transmission", y = "Miles per (U.S.) gallon")
```




```
as.data.frame(summarise(group_by(cars, Trans),
                        AverageMPG = round(mean(MPG), 1),
                        MedianMPG = median(MPG)))
```

```
##   Trans AverageMPG MedianMPG
## 1  Auto       17.1      17.3
## 2  Manu       24.4      22.8
```

The boxplot below compares cars with **automatic and manual transmission** according to their **MPG**. As shown in the plot, cars with **manual transmission yield a higher mileage than cars with an automatic transmission** regardless of any other for any other variable. The measures of central tendency, average and median, quantify this difference for the two groups of cars.

Initial (Unadjusted) Regression Model

```
# Build first regression model
fit1 <- lm(MPG ~ Trans, cars)
# Show regression coeficients
coef(fit1)
```

```
## (Intercept)   TransManu
##   17.147368    7.244939
```

The coeficients of our first, linear regression model yield the same result as described above. **MPG** is the model's outcome variable and **Trans** is the only regressor for now; hence, the model is unadjusted for any other variable.

The **intercept** (coefficient) describes the **average MPG** of the first transmission type: **automatic being the reference group** of this model. That confirms the average MPG that was computed before for cars with automatic transmission. The **coefficient of the regressor** being the other transmission type is interpretted as the change in the MPG mean comparing those with **manual transmission (TransManu)** to those without, hence, with automatic transmission as the reference group. The regression coefficient is positive and describes the incremental mileage (in MPG) of manual cars compared to automatic cars; that's also the difference of the averages that was computed above.

Additional Models Adjusted for Other Variables

Omitting other model variables (regressors) can bias the estimation of the coefficients of interest (**Trans**). Therefore, in addition to the initial model, alternative multivariate regression models will be constructed. Given our coefficient of interest, covariate adjustment is used and multiple models are built by adding other regressors to probe the effect (of car **Trans** on **MPG**) to evaluate for robustness and to see how other covariates influence the effect. Modeling multivariate relationships is difficult. Therefore, we try to play around with the dataset to see how the inclusion or exclustion of another variable can change the analysis.

On the one end of the spectrum of model possibilities, our first model (**fit1**) included just **one regressor**: **Trans**. In contrast, another model (**fit0**) is constructed that includes **all other variables of the dataset** as regressors (a.k.a. covariates) except the **outcome variable MPG**, of course.

```
# Build basic model that includes all variables
fit0  <- lm(          MPG ~ . , cars)
# Re-build first regression model with a different function that's based on the basic model
fit1  <- update(fit0, MPG ~ Trans)
# Compare regression coeficients
coef(fit1); coef(fit0)
```

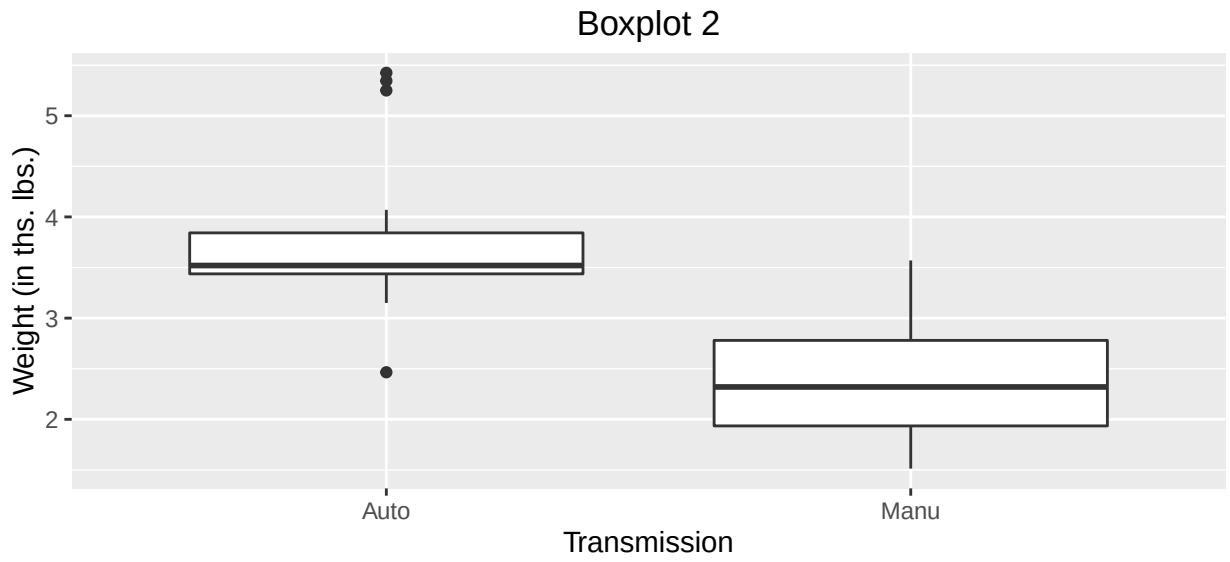
```
## (Intercept)   TransManu
##   17.147368    7.244939
```

```
## (Intercept)   TransManu      Weight         Cyl       Displ          HP
## 12.30337416  2.52022689 -3.71530393 -0.11144048  0.01333524 -0.02148212
##         RAR      QMTime         VS1        Gear        Carb
##  0.78711097  0.82104075  0.31776281  0.65541302 -0.19941925
```

In comparing the two pairs of coeficients (**Intercept and TransManu**) of the two models, the MPG averages for cars with **automatic and manual transmission** are lower once we adjust for all other covariates that may have an influence on **MPG**. Also, comparing the coeficients of the basic model (**fit0**) with one another, the average MPG of **manual** cars seems still higher compared to **automatic** cars, yet not as different as before when we compared without adjustment of other variables (**fit1**).

From a more statistics-, regression-unrelated and real-world point of view - although from the point of view of someone that has a limited knowledge of automobiles, yet a decent understanding of physics - , the **weight** of a car (here measured in ths. lbs.) seems to have a clear impact on its mileage. In other words, the heavier a car, the less miles it can reach with one gallon of fuel. The coeficient of the regressor **Weight** has the units **MPG per 1,000 lbs**. In other words, the coefficient of -3.7 is the expected change in miles per gallon for every aditional 1,000 lbs change in the weight of the car holding all of the other regressors fixed/constant.

```r
# Draw boxplot
ggplot(cars, aes(x = Trans, y = Weight)) + geom_boxplot() +
      labs(title = "Boxplot 2", x = "Transmission", y = "Weight (in ths. lbs.)")
```



As mentioned before, omitting other regressor (such as **weight**) can bias the estimation of the coefficient of certain other correlated regressors. **Boxplot 2** shows that automatic cars tend to be heavier than manual cars. So, what if the **effect of transmission on MPG** is due to the fact that car **transmission is correlated with weight**.

We're going to form a nested sequence of models. This means that we're going to add certain variables to the model, such as **weight** whereby the regressors of one model (**Trans**) are included in those of the next model.

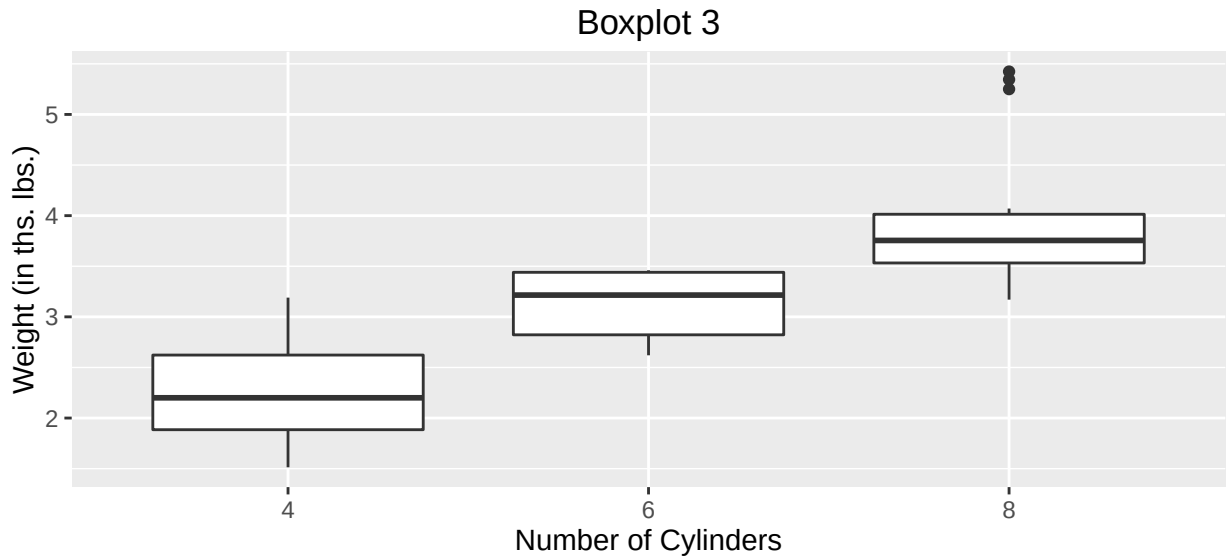
```r
# Center weight variable
cars  <- mutate(cars, Weight_centered = Weight-mean(Weight))
# Build nested regression model step 2
fit2  <- update(fit0, MPG ~ Trans + Weight_centered)
# Compare regression coeficients
coef(fit1); coef(fit2)
```

```
## (Intercept)   TransManu
##   17.147368    7.244939
```

```
##     (Intercept)       TransManu Weight_centered
##     20.10021868     -0.02361522     -5.35281145
```

Adjusting for the weight of a car, the regression **coeficient** (**fit2**) for **manual** transmissions has now **changed in magnitude and direction**. Accounting for weight, cars with manual transmission seem to have **less MPG on average than automatic** cars, although the difference is very marginal. The **intercept** (coeficient) is the expected MPG of an **automatic** car with an **average weight** since the other covariate, **weight**, was centered to its mean before included in the model. Without centering the weight, the intercept would have been interpreted as expected MPG of an **automatic** car with no weight, which is an unrealistic interpretation.

```
# Draw boxplot
ggplot(cars, aes(x = factor(Cyl), y = Weight)) + geom_boxplot() +
      labs(title = "Boxplot 3", x = "Number of Cylinders", y = "Weight (in ths. lbs.)")
```



Since the transmission of a car seems to be correlated with weight, what variables is weight correlated with? Again... From a layman's point of view, it is suggested that **cars increase in weight the more cylinders their engines carry**. **Boxplot 3** provides reason for this assumption.

The **Cyl** variable (number of cylinders) is therefore added to the model. Both number of cylinders and weight are centered so that the model intercept can be interpreted more realistically.

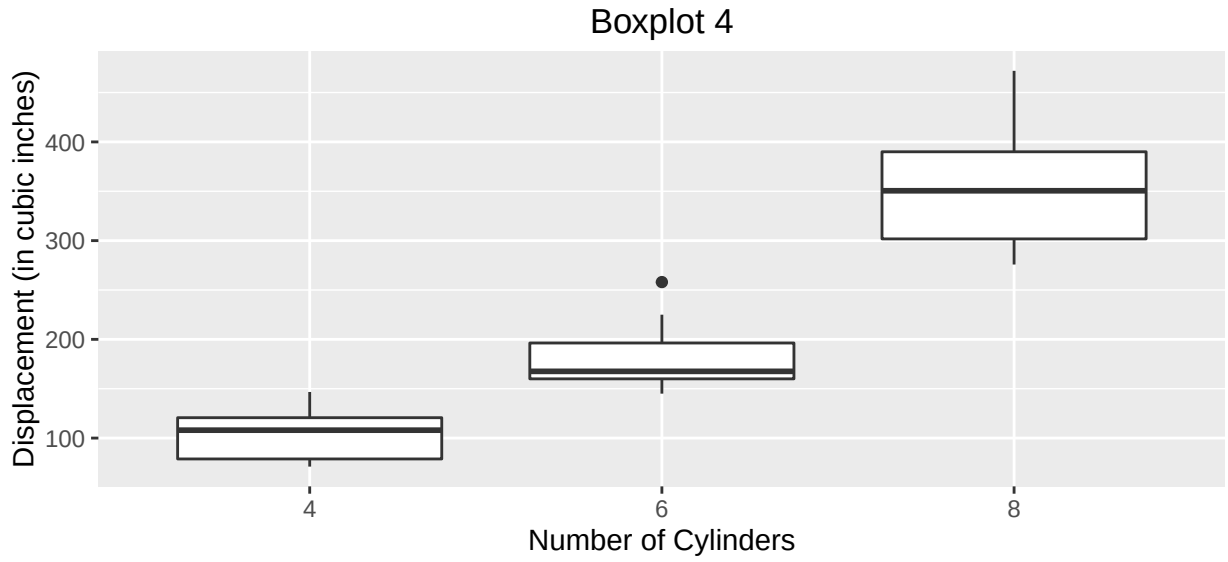
```
# Center cylinder no. variable
cars  <- mutate(cars, Cyl_centered = Cyl-mean(Cyl))
# Build nested regression model step 3
fit3  <- update(fit0, MPG ~ Trans + Weight_centered + Cyl_centered)
# Compare regression coeficients
coef(fit2);coef(fit3)
```

```
##     (Intercept)       TransManu Weight_centered
##     20.10021868     -0.02361522     -5.35281145

##     (Intercept)       TransManu Weight_centered    Cyl_centered
##      20.0189247       0.1764932      -3.1251422      -1.5102457
```

Adjusting for the number of cylinders and weight of a car, the regression coeficient for **manual** transmissions has changed again in direction, although the effect magnitude is comparably marginal as was the effect of the previous model. Accounting for the other two regressor variables, average-weight cars with an average number of cylinders and a **manual transmission seem to yield slightly more MPG on average than automatic** cars. The **intercept** (coeficient), the expected MPG of an **automatic** car with an average weight and an average number of cylinders, remains nearly the same.

```
# Draw boxplot
ggplot(cars, aes(x = factor(Cyl), y = Displ)) + geom_boxplot() +
      labs(title = "Boxplot 4", x = "Number of Cylinders", y = "Displacement (in cubic inches)")
```




```
# Compute correlation between weight and displacement of car
cor(cars$Weight, cars$Displ)
```

```
## [1] 0.8879799
```

```
# Center displacement variable
cars  <- mutate(cars, Displ_centered = Displ-mean(Displ))
# Build nested regression model step 4
fit4  <- update(fit0, MPG ~ Trans + Weight_centered + Cyl_centered + Displ_centered)
# Compare regression coeficients
coef(fit3); coef(fit4)
```

```
##     (Intercept)       TransManu Weight_centered    Cyl_centered
##      20.0189247       0.1764932      -3.1251422      -1.5102457

##     (Intercept)       TransManu Weight_centered    Cyl_centered
##    20.038192112     0.129065571    -3.583425472    -1.784173258
##  Displ_centered
##     0.007403833
```

Let's add another regressor to the model: Displacement (**Displ**). This variable is measured in cubic inches and describes – according to a layman's quick web search – "the volume of an engine's cylinders, a general indicator of its size and power". That relates well to the last two regressors (**weight and number of cylinders**) that had been added to the model as seen on **boxplot 4** and the correlation coeficient. The variable was centered as usual.

Adjusting for the number of cylinders, their volume (displacement), and the overall weight of a car, the **regression coeficients** (both intercept describing the **automatic** reference group and the one for the *manual transmission* factor level) have not changed significantly– neither in direction, nor in magnitude. Neither did the coeficients of the covariates, **weight and number of cylinders**. Holding all aforementioned covariates constant, the new regressor (**Displ**) doesn't seem to have an impact at all on MPG since its coeficient is nearly zero.

```r
# Center horsepower variable
cars  <- mutate(cars, HP_centered = HP-mean(HP))
# Build nested regression model step 5
fit5  <- update(fit0, MPG ~ Trans + Weight_centered + Cyl_centered + Displ_centered + HP_centered)
# Compare regression coeficients
coef(fit4); coef(fit5)
```

```
##     (Intercept)       TransManu Weight_centered    Cyl_centered
##    20.038192112     0.129065571    -3.583425472    -1.784173258
##  Displ_centered
##     0.007403833

##     (Intercept)       TransManu Weight_centered    Cyl_centered
##     19.45830027      1.55649163     -3.30262301     -1.10637984
##  Displ_centered     HP_centered
##      0.01225708     -0.02796002
```

One could add more and more covariates to the model and determine how adjustment for other variables affects the impact of a transmission on mileage (MPG). For instance, **the more cylinders a car's engine has, the more horsepower it yields**. (My father-in-law taught me that actually quite straight-forward lesson. It hadn't appeared to me before.) Hence, we expect that horsepower has a reason to be part of the model, as well. Holding all other covariates constant, the new regressor (**HP**) doesn't seem to have an impact at all on MPG since its coeficient is nearly zero– equivalent to the previously added regressor (**Displ**).

In general, the selection of covariates (their addition or removal) is a tricky endeavor. It depends heavily on how rich of a covariate space one wants to explore; the space of models explodes quickly as one adds interactions of variables and polynomial terms. We could have gone one and on until we eventually reach the basic multivariate regression model (fit0) that includes all regressors available. We could have gone beyond and investigate variable interactions. But for reasons of simplicity, we leave it with the current nested models and test them.

```r
# Compare nested models more holistically
anova(fit1, fit2, fit3, fit4, fit5)
```

```
## Analysis of Variance Table
##
## Model 1: MPG ~ Trans
## Model 2: MPG ~ Trans + Weight_centered
## Model 3: MPG ~ Trans + Weight_centered + Cyl_centered
## Model 4: MPG ~ Trans + Weight_centered + Cyl_centered + Displ_centered
## Model 5: MPG ~ Trans + Weight_centered + Cyl_centered + Displ_centered +
##     HP_centered
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.5432 7.017e-09 ***
## 3     28 191.05  1     87.27 13.9106 0.0009423 ***
## 4     27 188.43  1      2.62  0.4178 0.5236992
## 5     26 163.12  1     25.31  4.0336 0.0550966 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Including more regressors will reduce a model's **residual sum of squares (RSS)**, as seen in the **Analysis of Variance (ANOVA)** table, by the **Sum of Sq.** amounts. When adding regressors, the reduction in

residual sums of squares should be tested for significance above and beyond that of **reducing residual degrees of freedom (Res. Df)**. R's ANOVA function uses an ___F-test (PR>F)___for this purpose. It appears that each model is a significant improvement on its predecessor until the addition of the **Displ** variable in **model 4** which is not significant and should therefore be removed from the model sequence. and the subsequent model 5 from our ling approach.

```
# Update model
fit4  <- update(fit0, MPG ~ Trans + Weight_centered + Cyl_centered + HP_centered)
# Compare nested models more holistically
anova(fit1, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: MPG ~ Trans
## Model 2: MPG ~ Trans + Weight_centered
## Model 3: MPG ~ Trans + Weight_centered + Cyl_centered
## Model 4: MPG ~ Trans + Weight_centered + Cyl_centered + HP_centered
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 70.2925  5.39e-09 ***
## 3     28 191.05  1     87.27 13.8611 0.0009165 ***
## 4     27 170.00  1     21.05  3.3432 0.0785534 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We updated the nested models (**model 4**) and removed the *Displ* regressor and re-ran an ANOVA.

In a similar manner, the addition of the **horsepower (HP)** variable, now in the **updated model 4 and formerly in model 5**, is not significant and should therefore be removed from the nested models, as well. That brings us back to step 3 in our model nesting and leaves us with **model 3**. Since the addition of **weight** and number of cylinders (**Cyl**) **reduces the RSS significantly** (see RSS, Sum of Sq. and F-test p-values w/ significance codes in ANOVA table), **model 3** shall be the model to answer the overarching research question.

```
# Describe the whole more holistically
summary(fit3)
```

```
##
## Call:
## lm(formula = MPG ~ Trans + Weight_centered + Cyl_centered, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1735 -1.5340 -0.5386  1.5864  6.0812
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      20.0189     0.7029  28.481  < 2e-16 ***
## TransManu         0.1765     1.3045   0.135  0.89334
## Weight_centered  -3.1251     0.9109  -3.431  0.00189 **
## Cyl_centered     -1.5102     0.4223  -3.576  0.00129 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.612 on 28 degrees of freedom
## Multiple R-squared:  0.8303, Adjusted R-squared:  0.8122
## F-statistic: 45.68 on 3 and 28 DF,  p-value: 6.51e-11
```

```
# Compare R-squares
summary(fit1)$r.squared
```

```
## [1] 0.3597989
```

```
# Describe average weight (in 1,000 lbs) and number of cylinders of sampled cars
mean(cars$Weight);                          mean(cars$Cyl)
```

```
## [1] 3.21725
```

```
## [1] 6.1875
```

The model describes more than 80% of the total variation of a car's MPG as indicated by the **multiple and adjusted R-square** (a.k.a. regression variation). Adding more regressors to a model always increases the R-squares. Our initial model (**fit1**) with **Trans** being the only regressor explained a little more than a third of the **MPG** variance.

As seen in the coeficient table and previously interpreted, cars with a **manual transmission (TransManu)** yield slightly more mileage compared to its reference group being cars with **automatic transmission (intercept)** holding all other covariates fixed at their average measure (see above). The positive difference in mileage of cars with **manual vs. automatic transmission** is very marginal described by the value of the **TransManu coeficient** (less than 0.2 MPG). In addition, the effect is also **not significant indicated by the high p-value**. The coefficients in the table have **standard errors** and follow a **t (value)** distribution with n-p (number of cars, 32, minus number of regressors incl. the intercept, 4) **degrees of freedom**. Therefore, a t-interval hypothesis test of the estimated effect of a manual car compared to an automatic car can **quantify the uncertainty** of our estimate.

```
# Get a confidence interval for estimated MPG effect of car transmission
summary(fit3)$coefficients[2,1] + c(-1, 1) * qt(.975, df = fit3$df) * summary(fit3)$coefficients[2, 2]
```

```
## [1] -2.495555  2.848541
```

The computation of a confidence interval reveals that with 95% confidence, we estimate that a car with manual transmission can neither clearly increase nor decrease the MPG. The change in MPG ranges from -2.5 to 2.8 miles per gallon. In other words, we cannot confidently say what effect the transmission of car has on its mileage in MPG, if an effect at all.

```
# Re-visit regression-coefficient table
summary(fit3)$coef
```

```
##                   Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)     20.0189247  0.7028880 28.4809584 3.218877e-22
## TransManu        0.1764932  1.3044515  0.1353007 8.933421e-01
## Weight_centered -3.1251422  0.9108827 -3.4308942 1.885894e-03
## Cyl_centered    -1.5102457  0.4222792 -3.5764148 1.291605e-03
```

```r
# Get a confidence interval for estimated MPG effect of weight
summary(fit3)$coefficients[3,1] + c(-1, 1) * qt(.975, df = fit3$df) * summary(fit3)$coefficients[3, 2]
```

```
## [1] -4.991001 -1.259284
```

```r
# Get a confidence interval for estimated MPG effect of number of cylinders
summary(fit3)$coefficients[4,1] + c(-1, 1) * qt(.975, df = fit3$df) * summary(fit3)$coefficients[4, 2]
```

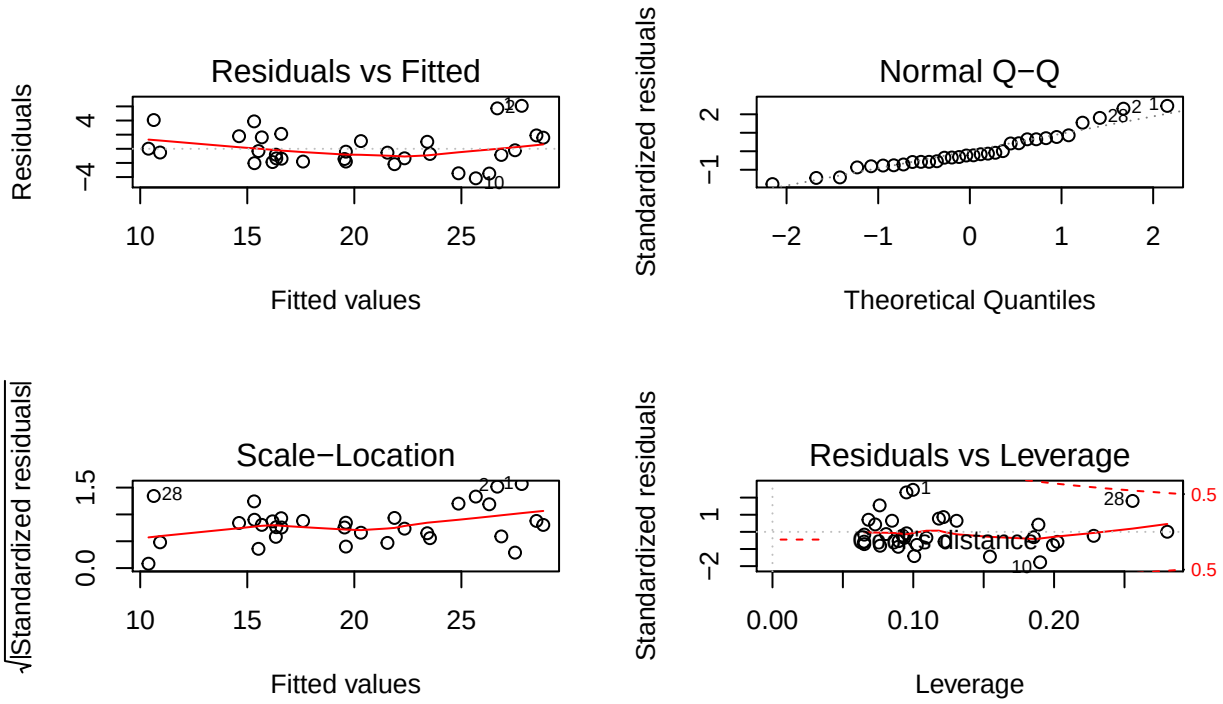
```
## [1] -2.3752454 -0.6452459
```

However, what we can say with **strong statistical significance** (due to p-values less than 1%) and **high confidence** (95%), as seen in the re-visited table of coeficients and confidence-interval computation, is that - every additional 1,000 lbs. in car weight descreases its mileage on average by 3.1 MPG; with 95% confidence, we estimate that a car with additional 1,000 lbs in weight can decrease the MPG ranging from at least 1.3 and up to 5 miles per gallon; - every additional cylinder in the engine of a car descreases its mileage on average by 1.5 MPG; with 95% confidence, we estimate that a car with every additional engine cylinder can decrease the MPG of the car ranging from at least .6 up to 2.4 miles per gallon.

That's being said with its ___covariates held constant at an average number of cylinders, respecively average weight, and with an automatic transmission (since that's the reference group of the transmission variable).

Model Diagnostics

```r
# Draw diagnostic plots
par(mfrow = c(2, 2)); plot(fit3)
```



```r
# Test residuals for normality
shapiro.test(resid(fit3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(fit3)
## W = 0.93688, p-value = 0.06108
```

Let's take a look at some diagnostic measures and plots, more specifically at the (top-left) **residual plot**. Residuals and residual plots are useful for investigating poor model fit. Positive residuals are above the line, negative residuals are below. Residuals can be thought of as the outcome (MPG) with the linear association of the predictors (all regressors of the model) removed. Residual plots highlight poor model fit, respectively in our case, acceptable model fit considering that most of the residuals are close to the zero line. The (top-right) **Residual QQ plot** investigates **normality of the errors**, an **assumption that's crucial to the ANOVA** we conducted earlier. The standardized residuals align fairly close to the diagonal line / theoretical quatiles of a normal distribution. The **Shapiro-Wilk test for normality** quantifies the observation from the plot and confirms the normal distribution of the model residuals. Model residuals are tested for normality to insure that the ANOVA applied.

Conclusions

A nested regression modelling approach was applied to determine the effect of a car's transmission type to its MPG. Initially, a regression model with just the transmission type as its predictor suggested a positive effect on MPG. However, once more car features were gradually added to the sequence of models the initial transmission effect has faded away. It turned out that weight of a car and its number of engine cylinders are more impactful on a car's MPG. Adjusting for other car features, one can say that neither an automatic or manual transmission is better for MPG. The MPG difference between automatic and manual transmissions is close to zero and can range slightly above or below zero depending on other variables included in the regression model.

For future analyses, it is suggested that instead of adding more and more variables to a nested sequence of models, one should start with the basic regression model (including all available regressors) and gradually removing car features that don't show an impact on MPG.