Regression Analysis: Motor Trend

Marcio Gualtieri

Contents

Required Packages

You might need to install and load the following packages if you don't already have them:

```
suppressMessages(library(xtable))                          # Pretty printing dataframes
suppressMessages(library(ggplot2))                         # Plotting
suppressMessages(library(gridExtra, warn.conflicts = FALSE))
suppressMessages(library(reshape2))                        # Transforming Data Frames
```

Overview

> ***Note: I don't think that makes any sense to move the supporting R code and figures to an "Appendix" section, thus such supporting items are placed where they are needed. Also, makes more sense to put the executive summary at the end (so one can use embedded R variables instead of typing values).***

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:

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

Data Loading

Our data-set consists of `mtcars`, which comes with R. You will find more information about it here.

```
data(mtcars)
```

Exploratory Data Analysis

Variables

R's documentation for this data-set provides some of the following information:

Variable	Description	Comment
mpg	Miles per (US) gallon	
cyl	Number of cylinders	
disp	Displacement (cu.in.)	Volume displaced when all engine's pistons perform a single move. Greater volume, greater engine.
hp	Gross horsepower	
drat	Rear axle ratio	Greater ratio, greater engine's RPM required to keep the same speed.
wt	Weight (1000 lbs)	
qsec	1/4 mile time	Least amount of time in seconds required to cover 1/4 of a mile. More powerful engine, least qsec.
vs	V/S (0 = V engine, 1 = Straight engine)	Pistons configuration in the engine, i.e., mounted in a "V" shape or straight.
am	Transmission (0 = automatic, 1 = manual)	
gear	Number of forward gears	In general manual transmission have more gears, but automatic has been catching up in recent years.

Variable	Description	Comment
carb	Number of carburetors	Cars these days use fuel injection, thus this makes me think this data-set is a bit old.

Data Types

You may inspect the schema by executiong the following command:

```
str(mtcars)
```

They are all numbers, but some should be categories. I will map them (`am` and `vs`,both categorical) to factors for easier reading:

```
mtcars$am <- ifelse(mtcars$am == 0, "Automatic", "Manual")
mtcars$am <- as.factor(mtcars$am)
mtcars$vs <- ifelse(mtcars$vs == 0, "V-Engine", "Straight Engine")
mtcars$vs <- as.factor(mtcars$vs)
```

Sampling Data

```
render_table <- function(data) {
  options(xtable.comment = FALSE)
  print(xtable(data))
}

sample_data_frame <- function(data, size) {
  sample_index <- sample(1:nrow(data), size)
  return(data[sample_index, ])
}

render_table(sample_data_frame(mtcars, 6))
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Merc 240D	24.40	4.00	146.70	62.00	3.69	3.19	20.00	Straight Engine	Automatic	4.00	2.00
Pontiac Firebird	19.20	8.00	400.00	175.00	3.08	3.85	17.05	V-Engine	Automatic	3.00	2.00
Cadillac Fleetwood	10.40	8.00	472.00	205.00	2.93	5.25	17.98	V-Engine	Automatic	3.00	4.00
Camaro Z28	13.30	8.00	350.00	245.00	3.73	3.84	15.41	V-Engine	Automatic	3.00	4.00
Lotus Europa	30.40	4.00	95.10	113.00	3.77	1.51	16.90	Straight Engine	Manual	5.00	2.00
Mazda RX4	21.00	6.00	160.00	110.00	3.90	2.62	16.46	V-Engine	Manual	4.00	4.00

Data Visualization

Scatter Plot Matrix

You may plot it by executing the following command:

```
pairs(mtcars, panel = panel.smooth, main = "Motor Trend", col = "light blue")
```

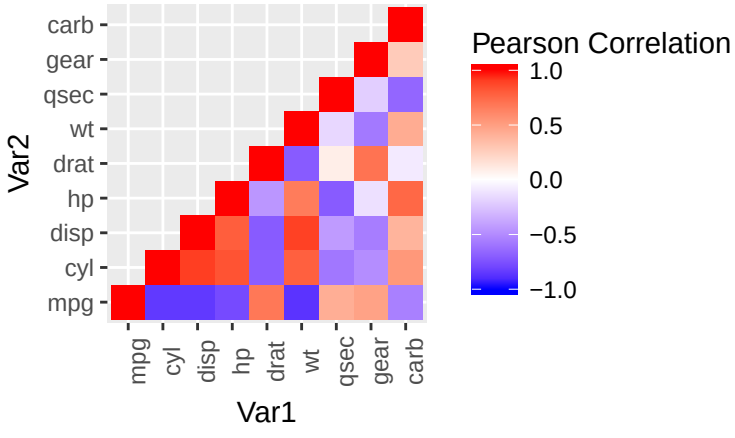
But I find correlation heat maps more compact and easier to read.

Correlation Heat Map

```
correlation_matrix <- function(data) {
  numeric_data <- data[, sapply(data,is.numeric)]
  matrix <- round(cor(numeric_data), 2)
  matrix[upper.tri(matrix)] <- NA
  matrix <- melt(matrix, na.rm = TRUE)
  return(matrix)
}

correlation_heat_map <- function(data) {
  matrix <- correlation_matrix(data)
  ggplot(data = matrix, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0, limit = c(-1, 1), name="Pearson Correlation") +
  theme(axis.text.x = element_text(angle = 90)) +
  coord_fixed()
}

correlation_heat_map(mtcars)
```



The potential predictors for MPG seem correlated among themselves to some degree, with a few exceptions:

- `gear` seems weakly correlated to `hp`.
- `drat` seems weakly correlated to `qsec` or `carb`.
- `wt` seems weakly correlated to `qsec`.

We can easily explain the pairs that correlate:

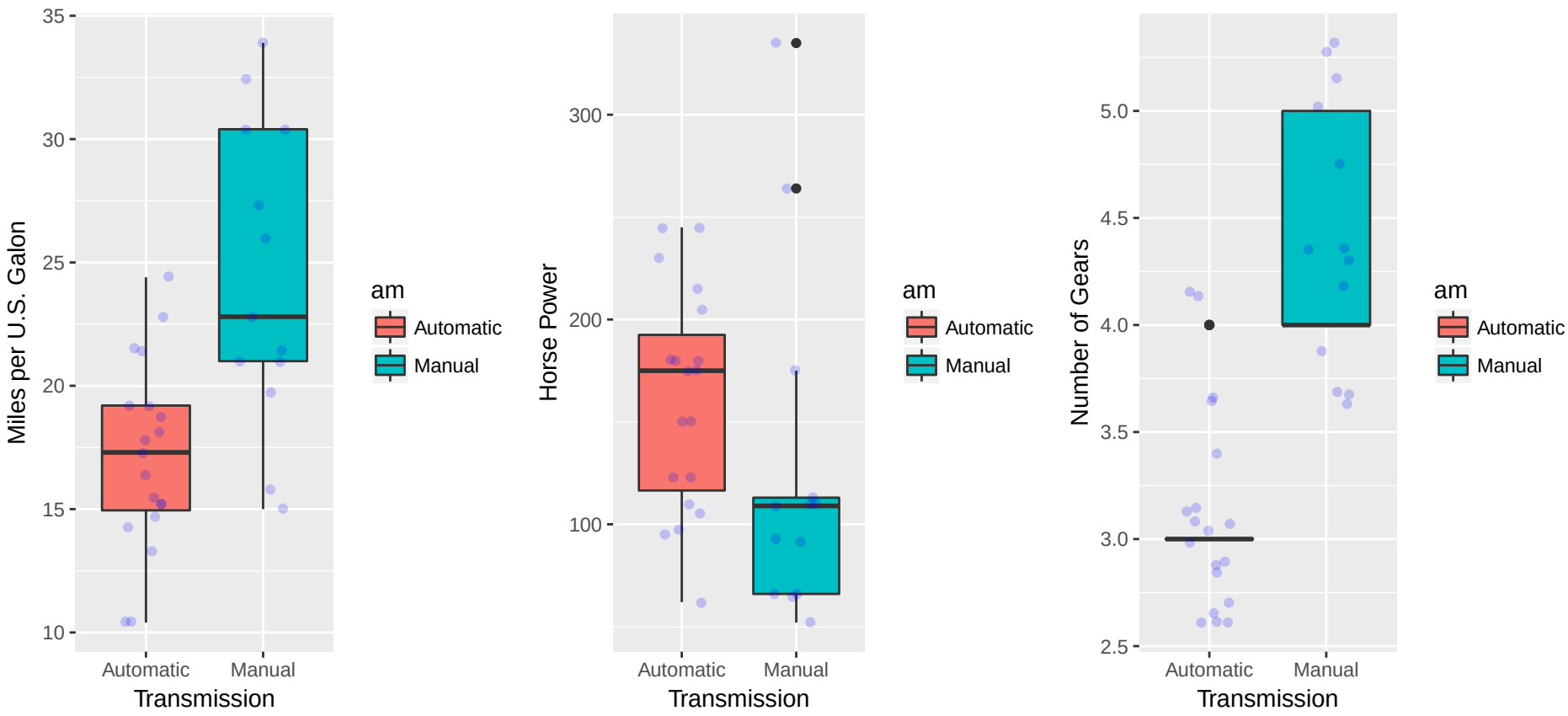
- Greater `wt` (weight), obviously implies more fuel consumption.
- Greater `cyl` (number of cylinders) `disp` (displacement volume) or `carb` (number of carburetors), implies more powerful engines and therefore more `hp` (horsepower). Greater `hp` (horsepower) generally implies less fuel efficiency (you need to go one way, efficiency, the other, power, or for a compromise between the two).
- Greater `gear` (number of gears) implies a greater number of degrees of freedom to choose the appropriate gear for a given speed, thus resulting in better fuel efficiency.
- `drat` (rear axle ratio) is a little trickier: greater the ratio, greater the engine's RPM (rotations per minute) required to keep the same speed, thus more fuel consumption.

Box Plots

Here are some box plots for the data:

```
box_plot <- function(data, y_column, x_column, x_title, y_title) {
  ggplot(data, aes_string(y = y_column, x = x_column)) +
  geom_boxplot(aes_string(fill = x_column)) +
  geom_point(position = position_jitter(width = 0.2), color = "blue", alpha = 0.2) +
  xlab(x_title) +
  ylab(y_title)
}

mpg_box <- box_plot(mtcars, "mpg", "am", "Transmission", "Miles per U.S. Galon")
hp_box <- box_plot(mtcars, "hp", "am", "Transmission", "Horse Power")
gear_box <- box_plot(mtcars, "gear", "am", "Transmission", "Number of Gears")
grid.arrange(mpg_box, hp_box, gear_box, ncol = 3)
```



From the box plots, we seem to have indeed better fuel efficiency for vehicles with automatic transmission for the following reasons:

- The vehicles with automatic transmission in the data-set seem to have greater horsepower, which correlates to less fuel efficiency.
- The vehicles with manual transmission have a greater number of gears.

Predictors Analysis

Linear Models

Let's try different models and look at their p-values to check their effect in the response (`mpg`):

```
summary(lm(mpg ~ . - 1, data = mtcars))$coef
```

```
##                       Estimate  Std. Error    t value   Pr(>|t|)
## cyl                -0.11144048  1.04502336 -0.1066392 0.91608738
## disp                0.01333524  0.01785750  0.7467585 0.46348865
## hp                 -0.02148212  0.02176858 -0.9868407 0.33495531
## drat                0.78711097  1.63537307  0.4813036 0.63527790
## wt                 -3.71530393  1.89441430 -1.9611887 0.06325215
## qsec                0.82104075  0.73084480  1.1234133 0.27394127
## vsStraight Engine 12.62113697 19.02841514  0.6632784 0.51436802
## vsV-Engine        12.30337416 18.71788443  0.6573058 0.51812440
## amManual           2.52022689  2.05665055  1.2254035 0.23398971
## gear               0.65541302  1.49325996  0.4389142 0.66520643
## carb              -0.19941925  0.82875250 -0.2406258 0.81217871
```

If you look at the P-values, all variables (but `qsec`) accept the null hypothesis $variable = 0$. The reason for that is that many of these variables correlate among themselves. For instance, the following predictors are obviously correlated:

```r
summary(lm(mpg ~ cyl + carb + disp + hp - 1, data = mtcars))$coef
```

```
##         Estimate Std. Error    t value     Pr(>|t|)
## cyl   8.05102716 1.08883280  7.3941813 4.722706e-08
## carb -1.24236370 1.57941786 -0.7865960 4.381256e-01
## disp -0.10883620 0.02604514 -4.1787525 2.597453e-04
## hp   -0.01769896 0.05625941 -0.3145955 7.554011e-01
```

The null hypothesis that $hp = 0$ is accepted due to its p-value. Using common sense and domain knowledge (refer to the `Correlation Heat Map` section), we may come up with the following model:

```r
fit_common_sense_with_gear <-lm(mpg ~ gear + hp + wt + drat, data = mtcars)
```

But given that the questions wee need to answer are related to transmission, let's replace `gear` with `am`, even though we know that `gear` is actually directly correlated to `mpg` and `am` is directly correlated to `gear`:

```r
fit_common_sense_with_am <- lm(mpg ~ am + hp + wt + drat, data = mtcars)
summary(fit_common_sense_with_am)
```

```
##
## Call:
## lm(formula = mpg ~ am + hp + wt + drat, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2882 -1.7531 -0.6827  1.1691  5.5211
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.027077   6.185177   4.855  4.5e-05 ***
## amManual     1.578521   1.559281   1.012 0.320363
## hp          -0.036373   0.009814  -3.706 0.000958 ***
## wt          -2.726092   0.937791  -2.907 0.007209 **
## drat         0.981018   1.377101   0.712 0.482341
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.56 on 27 degrees of freedom
## Multiple R-squared:  0.8428, Adjusted R-squared:  0.8196
## F-statistic:  36.2 on 4 and 27 DF,  p-value: 1.75e-10
```

We may automate this process using step as follows:

```r
fit_all <- lm(mpg ~ ., data = mtcars)
fit_best <- step(fit_all, direction = "both", trace = FALSE)
summary(fit_best)
```

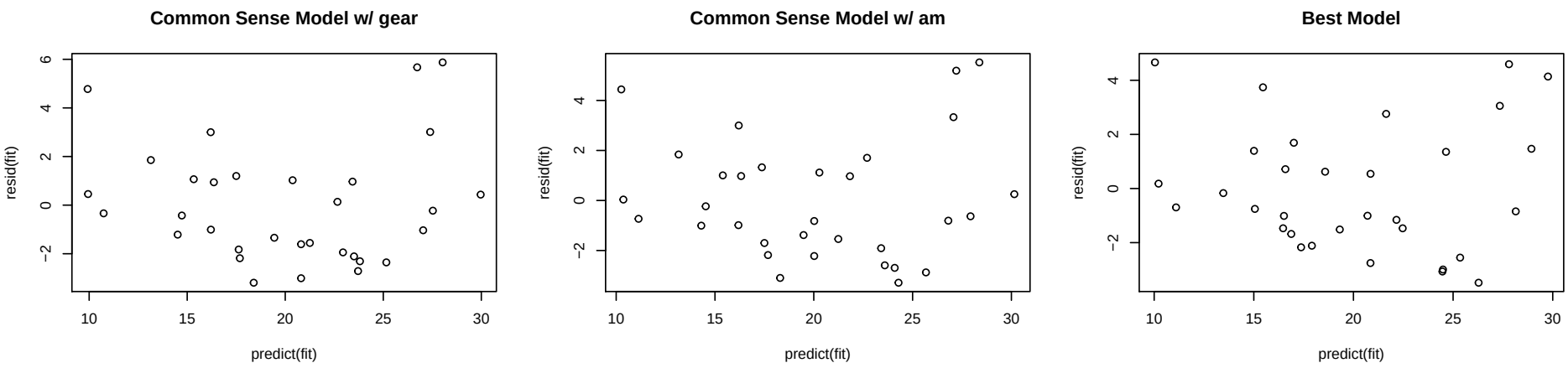
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

The "common sense" models did slightly worst in the residuals squared, but they are easier to understand.

Residuals

```r
residual_plot <- function(fit, title) plot(predict(fit), resid(fit), main = title)

par(mfrow = c(1, 3))
residual_plot(fit_common_sense_with_gear, "Common Sense Model w/ gear")
residual_plot(fit_common_sense_with_am, "Common Sense Model w/ am")
residual_plot(fit_best, "Best Model")
```



The residuals seem more or less randomly spread, thus uncorrelated to the response. This means our model is able to explain most of the behavior of the response.

Is an Automatic or Manual Transmission Better for MPG?

Statistical Inference

To answer this question, we only need a simple statistical inference using hypothesis test:

```
automatic <- mtcars[mtcars$am == "Automatic", ]
manual <- mtcars[mtcars$am == "Manual", ]
t.test(manual$mpg, automatic$mpg, alternative = "greater",
       paired = FALSE, var.equal = FALSE, conf.level = 0.95)$p.value
```

```
## [1] 0.0006868192
```

Conclusion

Given this p-value, which is less than $\alpha = 0.05$, we reject the null hypothesis. We know with 95% confidence that MPG for the manual transmission has greater MPG than the automatic transmission.

Quantify the MPG Difference Between Automatic and Manual Transmissions

Statistical Regression

We will use statistical regression to quantify this difference:

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

Conclusion

Let's take the intercept and the slope for the unadjusted estimate:

```
intercept_am <- coefficients(fit_am)[1]
slope_am <- coefficients(fit_am)[2]
r_squared_am <- summary(fit_am)$r.squared
```

The intercept (17.1473684) represents the mean `MPG` when `am` is zero (automatic). The slope (7.2449393) represents the increase in the mean MPG when `am` is one (manual), thus the mean MPG when `am` is one is $slope + intercept \times 1$, which is 24.3923077.

The above model only accounts for `am` and doesn't adjust for the effect of the other predictors. Therefore the slope itself, 7.2449393, doesn't quantify the difference between the `MPG` for automatic and manual transmissions (just look at R^2 of 0.3597989, which menas the model doesn't explain a lot of the data).

The following "common sense" model uses `am`, `hp`, `wt` and `drat`, therefore adjusting `am` for the effect of `hp`, `wt` and `drat`:

```
intercept_common_sense_with_am <- coefficients(fit_common_sense_with_am)[1]
slope_common_sense_with_am <- coefficients(fit_common_sense_with_am)[2]
r_squared_common_sense_with_am <- summary(fit_common_sense_with_am)$r.squared
```

The R^2 of 0.8428442 shows that this model explains the data much better.

The intercept doesn't have a physical interpretation here, since it would be the MPG for `amAutomatic` when the remaining predictors are zero (zero horsepower, weight and drat don't make much sense in an experimental setup). But the slope 1.5785208 represents the increase in MPG when switching from `amAutomatic` to `amManual` while keeping the remaining predictors constant.

Executive Summary

- Counter-intuitively, manual transmissions are more fuel efficient. One would think that automatic transmissions would change gears more efficiently, resulting in fuel savings, but that's not the case. The transmission type is indirectly correlated to MPG through the number of gears. In general, automatic transmissions have less gears than their manual counterparts. Greater number of gears means greater number of degrees of freedom when choosing the proper gear for each speed, thus resulting in fuel savings. Of course, we are assuming that the drivers used in the experiment that collected the data were experienced professionals, which can change gears very efficiently. Switching to manual transmission increases `mpg` (miles per U.S. galon) in 1.5785208 in average.
- Horsepower and vehicle's weight affect fuel efficiency much more heavily than transmission type.
- Another variable correlated to fuel efficiency is the rear axle ratio, which determines the engine's RPM's (revolutions per minute) required to keep the same speed. Greater rear axle ratio results in greater engine's RPM's for the same speed, thus greater fuel consumption.