Peer-graded Assignment: Regression Models Course Project

Introduction: 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

1 Load the mtcars data and perform some basic exploratory data analyses.

According to the description, the data contains 32 observations on 11 (numeric) variables.

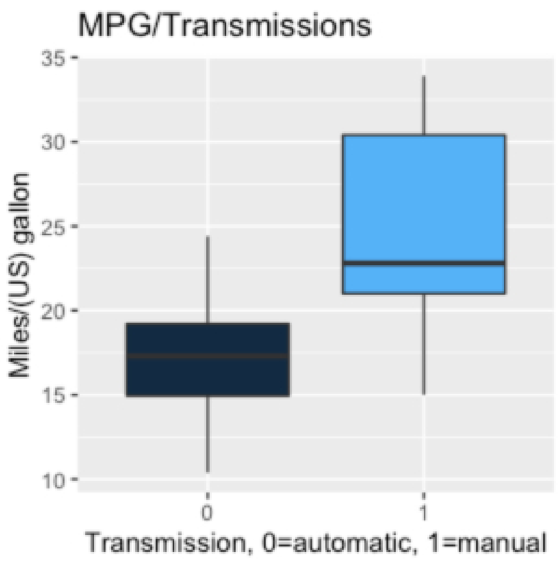
1. MPG Miles/(US) gallon
2. cyl Number of cylinders
3. disp Displacement (cu.in.)
4. hp Gross horsepower
5. drat Rear axle ratio
6. wt Weight (1000 lbs)
7. qsec 1/4 mile time
8. vs Engine (0 = V-shaped, 1 = straight)
9. am Transmission (0 = automatic, 1 = manual)
10. gear Number of forward gears
11. carb Number of carburetors

The data structure is:

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

2. Provide a basic summary of the MPG verus transmissions (automatic and

2. Provide a basic summary of the MPG verus transmissions (automatic and manual)



From this plot, it appeared that a manual transmission was better than an automatic transmission for MPG. More basic summaries in Appendix A also suggested that the manual transmission was better.

However, in the above plot, transmission was not adjusted for other terms. From the multivariable comparison chart in Appendix A, MPG was obviously correlated with many terms.

So, we wanted fit the data with a model with only necessary terms.

3. Models Selection

From the multivariable comparison chart in the Appendix A, we saw that variables were correlated. So, MPG might not be affected only by transmission(am), but also by some other factors.

To avoid unnecessary terms and include only necessary terms in our model, we applied a backward selection approach to fit our model. This method slowly removed one factor at a time, starting with the term with the highest p-value.

We took out the term with the highest p-value if its p-value were higher than above a specified p-value threshold (5%). We updated the model and checked the next term with the highest P-value. This continued until all the remaining terms in the model were below a specified p-value threshold.

```
# Initial our model with all terms
fit <- lm(mpg ~ . , data = mtcars)
# Show initial P-values
summary(fit)$coef
```

```
##                Estimate  Std. Error    t value   Pr(>|t|)
## (Intercept) 12.30337416 18.71788443  0.6573058 0.51812440
## 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
## vs           0.31776281  2.10450861  0.1509915 0.88142347
## am           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
```

```
# Start the backward selection approach
require(MASS)
step <- stepAIC(fit, direction="backward")
```

Steps can be found in the Appendix B.

```
step$anova # display results
```

```
## Stepwise Model Path 
## Analysis of Deviance Table
## 
## Initial Model:
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## 
## Final Model:
## mpg ~ wt + qsec + am
## 
## 
##     Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                             21   147.4944 70.89774
## 2  - cyl  1 0.07987121        22   147.5743 68.91507
## 3   - vs  1 0.26852280        23   147.8428 66.97324
## 4 - carb  1 0.68546077        24   148.5283 65.12126
## 5 - gear  1 1.56497053        25   150.0933 63.45667
## 6 - drat  1 3.34455117        26   153.4378 62.16190
## 7 - disp  1 6.62865369        27   160.0665 61.51530
## 8   - hp  1 9.21946935        28   169.2859 61.30730
```

```
# Update our model
fit_combined <- update(fit, mpg ~  wt + qsec + factor(am), data = mtcars)
# Show final P-values
# Show final P-values
summary(fit_combined)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + factor(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 ***
## factor(am)1   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
```

As a result, all P-values were smaller than 5%. So, according to our criterion, we could reject there were no difference between our terms and MPG, which suggested that these three interaction terms were necessary.

Since factor(am)1 was maller than 5%, it implied there was difference for MPG between automatic and manual transmission.

Besides, the adjusted R-squared was 0.8336, meaning that 83.36% of the variance of the MPG could be explained by this model.

Model Adjustment

a. Model with factor variables

Since we wanted to know if there was any difference between automatic and manual transmission to MPG, and the model already included them with all necessary terms.

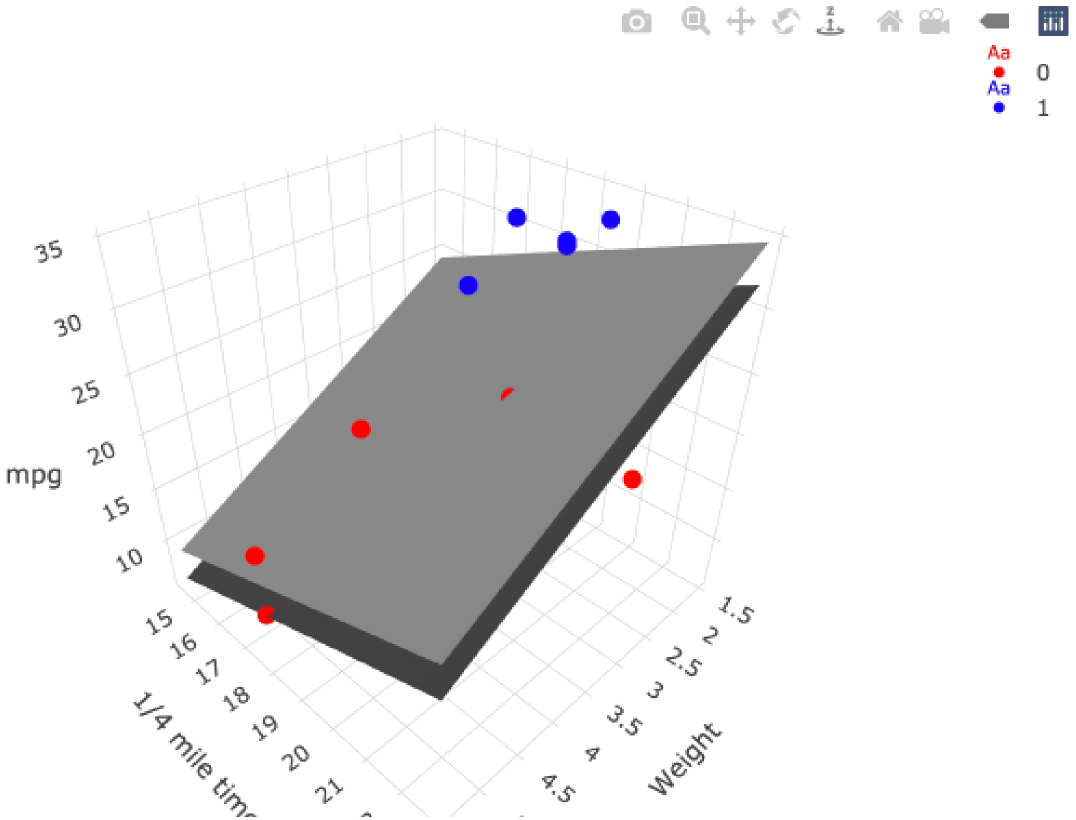
We saw that in addition to the intercept(referred to MPG), and slopes for “wt” and “qsec”, there was a third variable ‘factor(am)1’. When lm() encountered a factor variable with two levels, it created a new variable based on the second level.

In our case, the term “am” was a binary variable that took the value 1 if the transmission was manual, and 0 if it was automatic, and therefore ‘factor(am)1’ is created. The fitted equation for two groups can be written as

```
Automatic transmission: mpg = 9.6178 -3.9165(wt)  + 1.2259(qsec)
Manual transmission: mpg = 9.6178 + 2.9358 * 1 -3.9165(wt)  + 1.2259(qsec)
```

As a result, if the transmission is manual, we will consider the term “2.9358(manual)”, otherwise, we will ignore the term “2.9358(manual)”.

We can see that the slopes for “wt and”qsec" are the same. The intercept for automatic transmission is 9.6178. The intercept for a manual transmission is 9.6178 + (2.9358 * 1) = 12.5536.



(0 = automatic(light gray), 1 = manual(dark gray))

If we visualized the data (Weight, mpg, qsec) and fitted a linear model for two groups, (red = automatic; blue = manual), the regression plane (light gray) for the manual transmission positioned higher that the the regression plane (dark gray) for the automatic transmission in terms of mpg.

- Is an automatic or manual transmission better for miles/(US) gallon (MPG) ?
- Quantifying how different is the MPG between automatic and manual transmissions?

Since the intercept represents the mean value for MPG, it implies that for the same input to the formula, a manual transmission is better for MPG on average. And it will be 2.9358 Miles/(US) gallon better.

b. Model with interaction

From the previous plot, we could see that the regression plane (light gray) for the manual transmission did not align well with the data (blue color). We might be able to adjust the model with an interaction term. It means to split the data into automatic and manual transmission and fit a linear model for each group.

```
# Split into two groups and fit two models
mtcars0 <- mtcars[mtcars$am==0,]
mtcars1 <- mtcars[mtcars$a==1,]
fit_am0 <- lm(mpg ~ wt + qsec, data = mtcars0)
summary(fit_am0)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec, data = mtcars0)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.118 -1.363 -1.011  1.181  3.800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  11.2489     6.7148   1.675 0.113316
## wt           -2.9963     0.6636  -4.515 0.000352 ***
## qsec          0.9454     0.2945   3.210 0.005464 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.033 on 16 degrees of freedom
## Multiple R-squared:  0.7502, Adjusted R-squared:  0.7189
## F-statistic: 24.02 on 2 and 16 DF,  p-value: 1.518e-05
```

```
fit_am1 <- lm(mpg ~ wt + qsec, data = mtcars1)
summary(fit_am1)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec, data = mtcars1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6832 -0.9178 -0.0709  0.5567  4.0907
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  20.1754    11.1991   1.802 0.101799
## wt           -6.7544     1.4306  -4.721 0.000815 ***
## qsec          1.1810     0.4925   2.398 0.037434 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.244 on 10 degrees of freedom
## Multiple R-squared:  0.8896, Adjusted R-squared:  0.8675
## F-statistic: 40.29 on 2 and 10 DF,  p-value: 1.639e-05
```

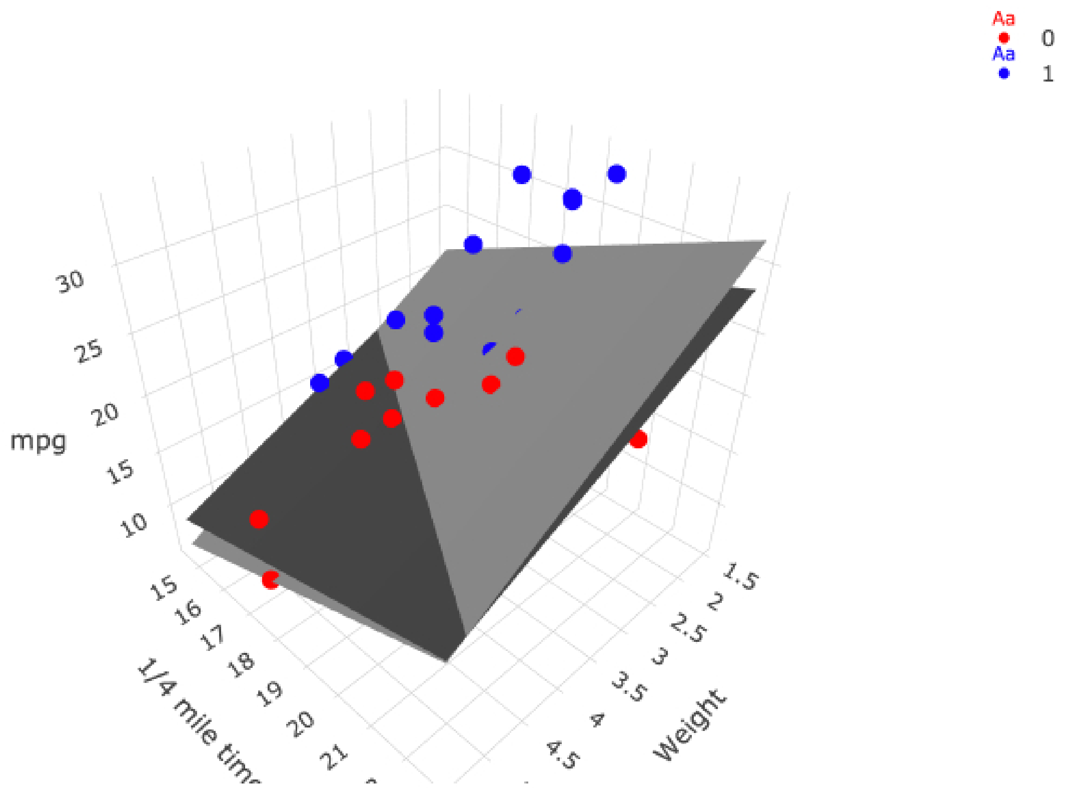
Because of the presence of the interaction term (transmission), both the slopes and intercepts are different for automatic and manual.

Also note that the coefficients of intercepts, having p-values highter than 0.05, are not statistically significant, meaning that the intercepts for automatic and manual are probably not very different. However, the slopes for automatic and manual are different.

The regression equation above can be split into separate equations for two groups:

`Automatic transmission: mpg = 11.2489 -2.9963(wt)  + 0.9454(qsec)`

`Manual transmission: mpg = 20.1754 -6.7544(wt)  + 1.1810(qsec)`



(0 = automatic(light gray), 1 = manual(dark gray))

The regression planes are not parallel, but the interaction term is very small and statistically insignificant, so if we could visualize them they would be almost parallel.

The model predicts that mpg increase with lower weight, but for the automatic transmission the slope of this increase is smaller. The slopes intersect at between around 5000 and 2653 lbs. For transmission with less than 2653 lbs weight, the mpg of manual is higher compared to the mpg of automatic transmission with the same details.

1/4 time mile has an additional positive effect on mpg. The slopes intersect at between around 14.5 and 22.9 lbs. Thus the maunal transmission with “1/4 mile time” greater than 14.5, will have higher mpg compared to automatic ones of the same details.

Nevertheless, the difference in slopes is due to the interaction term which is not statistically significant. So there is no evidence in the data that slopes should be different. No adjustment is needed.

4 Models diagnostics

We diagnosed our model with factor variables by the variance inflation factor(VIF). The VIF estimates how much the variance of a regression coefficient is inflated due to multicollinearity in the model. If VIF values were greater than 10, it indicated that terms were highly collinear with the other terms in the model.

```
library(car)
vif(fit_combined)
```

```
##        wt      qsec factor(am) 
##  2.482952  1.364339  2.541437
```

The result of VIF was satisfactory. They were all under 10. Besides, we also used diagnostic plots to provide checks for heteroscedasticity, normality, and influential observations. You may find them in the Appendix D. The plots showed that there was no a very clear systematic pattern in our residuals.

As a result, we were confident with our first model with factor variables.

Summary

With all the previous analysis, we can conclude that our regression model is a fit.

```
Automatic transmission: mpg = 9.6178 -3.9165(wt)  + 1.2259(qsec)
Manual transmission: mpg = 9.6178 + 2.9358 * 1 -3.9165(wt)  + 1.2259(qsec)
```

The weights and 1/4 mile time are related to mpg, but the adjusted estimate (mpg) depends on group status. It estimates that the manual transmission has a higher mpg than the automatic transmission.

Appendix A: Data exploration

```
# Summary for the automatic transmission
summary(mtcars[mtcars$am==0,])
```

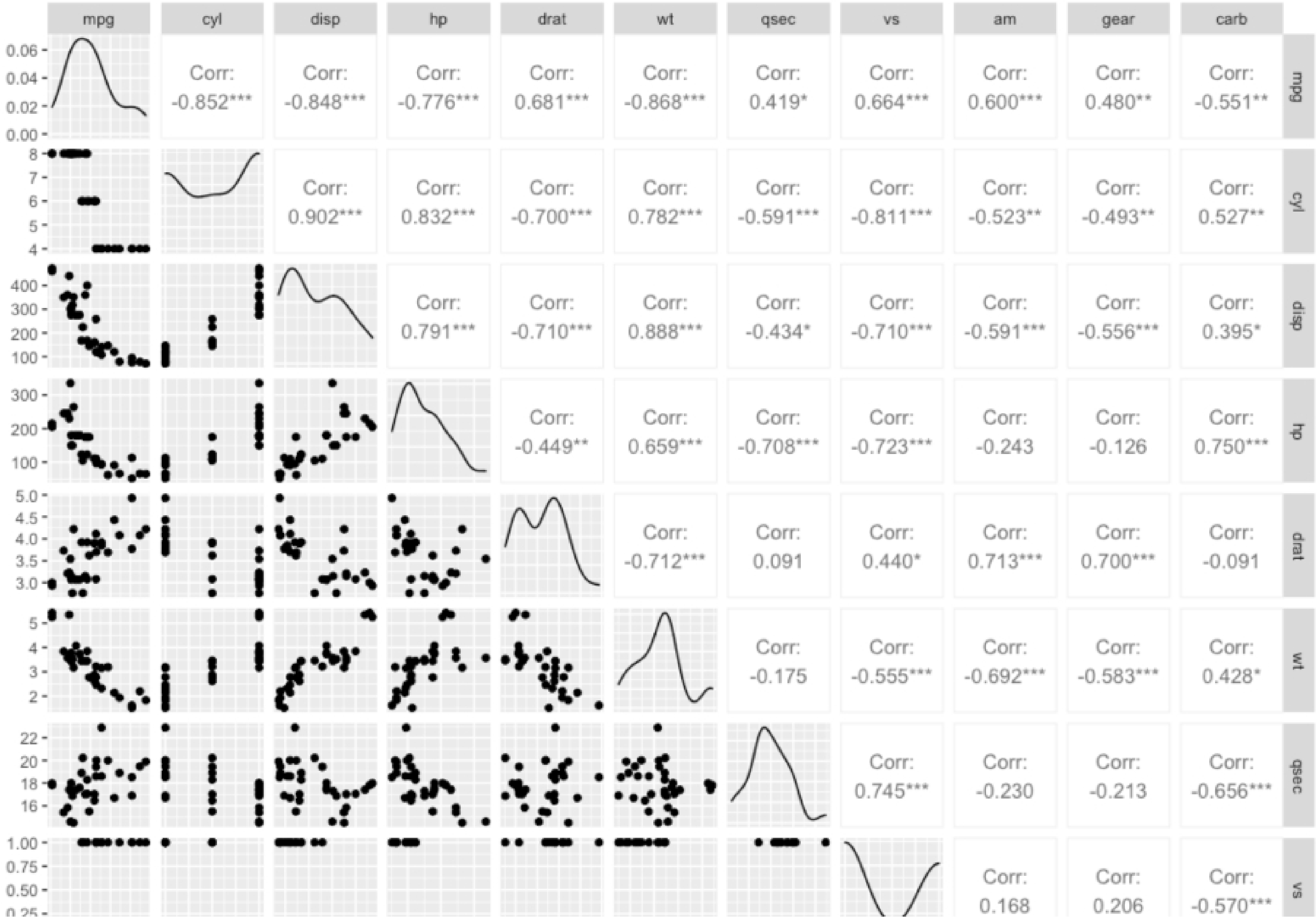
```r
summary(mtcars[mtcars$am==0,])
```

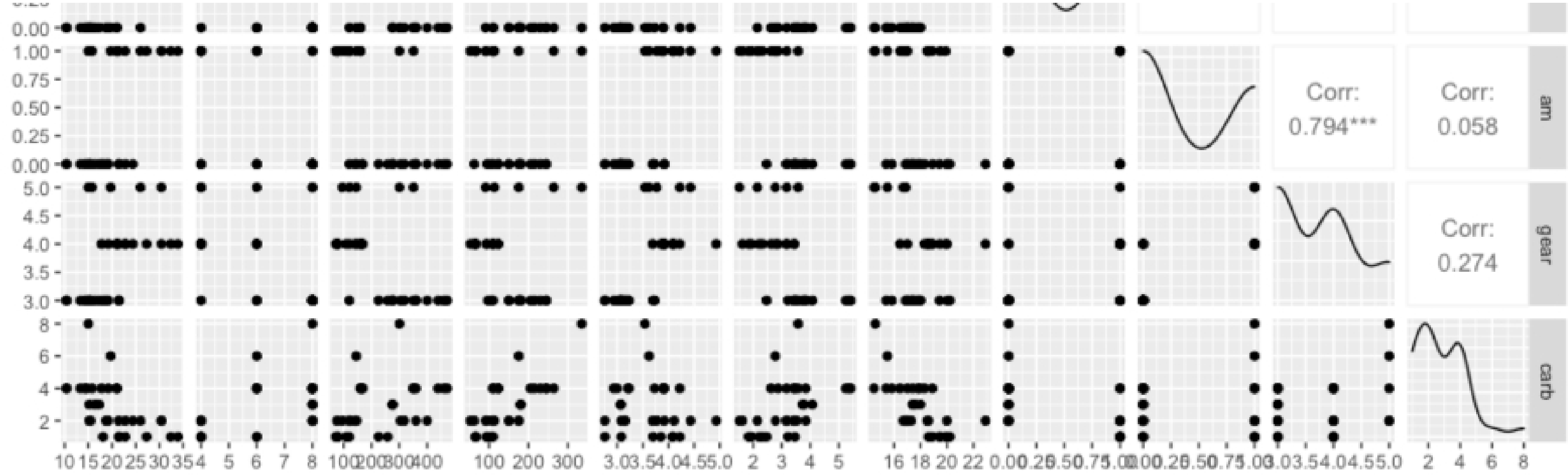
```
##       mpg             cyl             disp             hp       
##  Min.   :10.40   Min.   :4.000   Min.   :120.1   Min.   : 62.0  
##  1st Qu.:14.95   1st Qu.:6.000   1st Qu.:196.3   1st Qu.:116.5  
##  Median :17.30   Median :8.000   Median :275.8   Median :175.0  
##  Mean   :17.15   Mean   :6.947   Mean   :290.4   Mean   :160.3  
##  3rd Qu.:19.20   3rd Qu.:8.000   3rd Qu.:360.0   3rd Qu.:192.5  
##  Max.   :24.40   Max.   :8.000   Max.   :472.0   Max.   :245.0  
##       drat            wt             qsec             vs               am   
##  Min.   :2.760   Min.   :2.465   Min.   :15.41   Min.   :0.0000   Min.   :0  
##  1st Qu.:3.070   1st Qu.:3.438   1st Qu.:17.18   1st Qu.:0.0000   1st Qu.:0  
##  Median :3.150   Median :3.520   Median :17.82   Median :0.0000   Median :0  
##  Mean   :3.286   Mean   :3.769   Mean   :18.18   Mean   :0.3684   Mean   :0  
##  3rd Qu.:3.695   3rd Qu.:3.842   3rd Qu.:19.17   3rd Qu.:1.0000   3rd Qu.:0  
##  Max.   :3.920   Max.   :5.424   Max.   :22.90   Max.   :1.0000   Max.   :0  
##       gear            carb      
##  Min.   :3.000   Min.   :1.000  
##  1st Qu.:3.000   1st Qu.:2.000  
##  Median :3.000   Median :3.000  
##  Mean   :3.211   Mean   :2.737  
##  3rd Qu.:3.000   3rd Qu.:4.000  
##  Max.   :4.000   Max.   :4.000
```

```r
# Summary for the manual transmission
summary(mtcars[mtcars$am==1,])
```

```
##       mpg             cyl             disp             hp             drat     
##  Min.   :15.00   Min.   :4.000   Min.   : 71.1   Min.   : 52.0   Min.   :3.54  
##  1st Qu.:21.00   1st Qu.:4.000   1st Qu.: 79.0   1st Qu.: 66.0   1st Qu.:3.85  
##  Median :22.80   Median :4.000   Median :120.3   Median :109.0   Median :4.08  
##  Mean   :24.39   Mean   :5.077   Mean   :143.5   Mean   :126.8   Mean   :4.05  
##  3rd Qu.:30.40   3rd Qu.:6.000   3rd Qu.:160.0   3rd Qu.:113.0   3rd Qu.:4.22  
##  Max.   :33.90   Max.   :8.000   Max.   :351.0   Max.   :335.0   Max.   :4.93  
##        wt             qsec             vs               am            gear      
##  Min.   :1.513   Min.   :14.50   Min.   :0.0000   Min.   :1   Min.   :4.000  
##  1st Qu.:1.935   1st Qu.:16.46   1st Qu.:0.0000   1st Qu.:1   1st Qu.:4.000  
##  Median :2.320   Median :17.02   Median :1.0000   Median :1   Median :4.000  
##  Mean   :2.411   Mean   :17.36   Mean   :0.5385   Mean   :1   Mean   :4.385  
##  3rd Qu.:2.780   3rd Qu.:18.61   3rd Qu.:1.0000   3rd Qu.:1   3rd Qu.:5.000  
##  Max.   :3.570   Max.   :19.90   Max.   :1.0000   Max.   :1   Max.   :5.000  
##       carb      
##  Min.   :1.000  
##  1st Qu.:1.000
##  Median :2.000
##  Mean   :2.923
##  3rd Qu.:4.000
##  Max.   :8.000
```

```
require(GGally)
# Multivariable Comparison
g <- ggpairs(mtcars)
g
```





Appendix B: Model selection steps

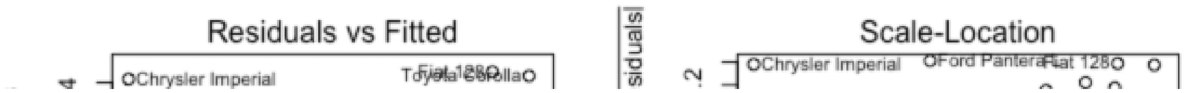
```
fit <- lm(mpg ~ . , data = mtcars)
# Start the backward selection approach
require(MASS)
step <- stepAIC(fit, direction="backward")
```

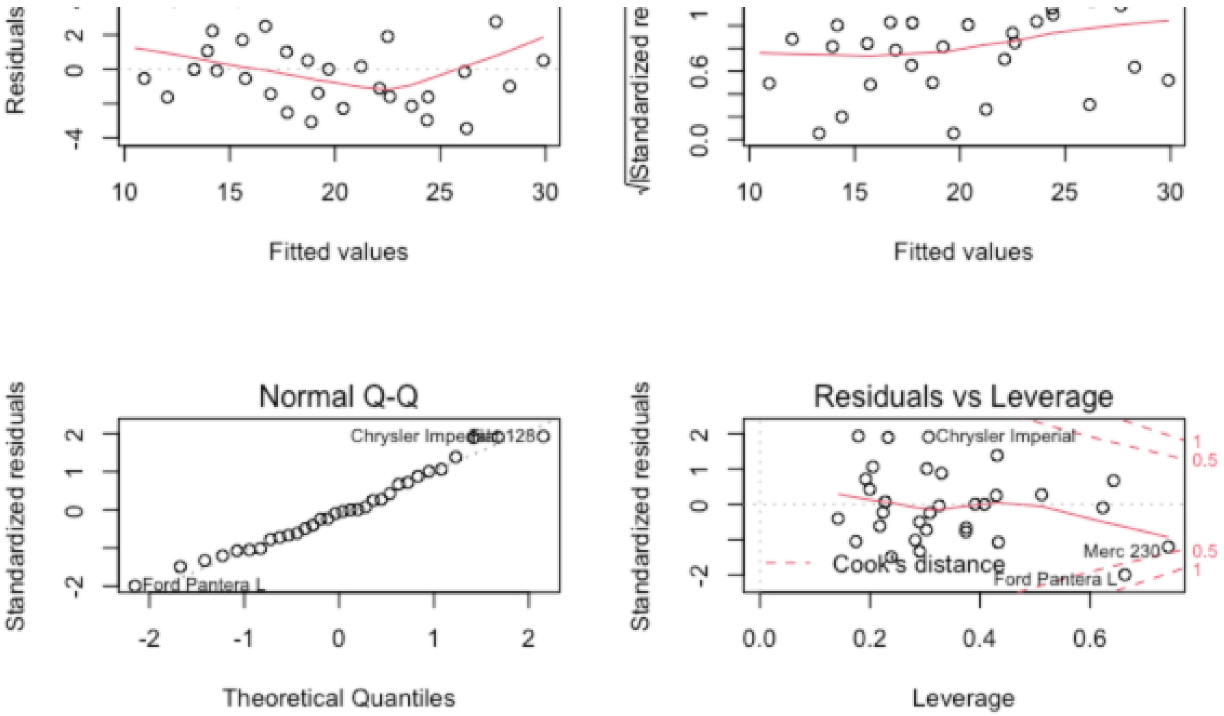
```
## Start:  AIC=70.9
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - cyl   1    0.0799 147.57 68.915
## - vs    1    0.1601 147.66 68.932
## - carb  1    0.4067 147.90 68.986
## - gear  1    1.3531 148.85 69.190
## - drat  1    1.6270 149.12 69.249
## - disp  1    3.9167 151.41 69.736
## - hp    1    6.8399 154.33 70.348
## - qsec  1    8.8641 156.36 70.765
## <none>              147.49 70.898
## - am    1   10.5467 158.04 71.108
## - wt    1   27.0144 174.51 74.280
##
## Step:  AIC=68.92
## mpg ~ disp + hp + drat + wt + qsec + vs + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - vs    1    0.2685 147.84 66.973
## - carb  1    0.5201 148.09 67.028
## - gear  1    1.8211 149.40 67.308
## - drat  1    1.9826 149.56 67.342
## - disp  1    3.9009 151.47 67.750
## - hp    1    7.3632 154.94 68.473
## <none>              147.57 68.915
## - qsec  1   10.0933 157.67 69.032
## - am    1   11.8359 159.41 69.384
## - wt    1   27.0280 174.60 72.297
##
## Step:  AIC=66.97
## mpg ~ disp + hp + drat + wt + qsec + am + gear + carb
##
##        Df Sum of Sq    RSS    AIC
## - carb  1    0.6855 148.53 65.121
## - gear  1    2.1437 149.99 65.434
## - drat  1    2.2139 150.06 65.449
## - disp  1    3.6467 151.49 65.753
## - hp    1    7.1060 154.95 66.475
## <none>              147.84 66.973
## - am    1   11.5694 159.41 67.384
## - qsec  1   15.6830 163.53 68.200
## - wt    1   27.3799 175.22 70.410
##
## Step:  AIC=65.12
## mpg ~ disp + hp + drat + wt + qsec + am + gear
##
##        Df Sum of Sq    RSS    AIC
## - gear  1     1.565 150.09 63.457
## - drat  1     1.932 150.46 63.535
## <none>              148.53 65.121
## - disp  1    10.110 158.64 65.229
## - am    1    12.323 160.85 65.672
## - hp    1    14.826 163.35 66.166
## - qsec  1    26.408 174.94 68.358
## - wt    1    69.127 217.66 75.350
##
## Step:  AIC=63.46
## mpg ~ disp + hp + drat + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - drat  1     3.345 153.44 62.162
## - disp  1     8.545 158.64 63.229
## <none>              150.09 63.457
## - hp    1    13.285 163.38 64.171
## - am    1    20.036 170.13 65.466
## - qsec  1    25.574 175.67 66.491
## - wt    1    67.572 217.66 73.351
##
## Step:  AIC=62.16
## mpg ~ disp + hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - disp  1     6.629 160.07 61.515
## <none>              153.44 62.162
## - hp    1    12.572 166.01 62.682
## - qsec  1    26.470 179.91 65.255
## - am    1    32.198 185.63 66.258
## - wt    1    69.043 222.48 72.051
##
## Step:  AIC=61.52
## mpg ~ hp + wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## - hp    1     9.219 169.29 61.307
## <none>              160.07 61.515
## - qsec  1    20.225 180.29 63.323
## - am    1    25.993 186.06 64.331
## - wt    1    78.494 238.56 72.284
##
## Step:  AIC=61.31
## mpg ~ wt + qsec + am
##
##        Df Sum of Sq    RSS    AIC
## <none>              169.29 61.307
## - am    1    26.178 195.46 63.908
## - qsec  1   109.034 278.32 75.217
## - wt    1   183.347 352.63 82.790
```

Appendix C: Models visualization

Model with factor variables `Automatic transmission: mpg = 9.6178 -3.9165(wt)  + 1.2259(qsec)` `Manual transmission: mpg = 9.6178 + 2.9358 * 1 -3.9165(wt)  + 1.2259(qsec)`

Appendix D: Diagnostic plots




Residuals
Fitted values
√|Standardized re
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Chrysler Imperial
Fiat 128
Ford Pantera L
Residuals vs Leverage
Standardized residuals
Leverage
Chrysler Imperial
Merc 230
Ford Pantera L
Cook's distance
1
0.5
0.5
1