Automatic vs Manual Transmission on MPG

Regression Models Course Project

Xinchun Tian

Executive Summary

High dimensional data analysis has become increasingly frequent and important in diverse fields of sciences, engineering, and humanities, ranging from genomics and health sciences to economics, finance and machine learning [1]. Generally the inclusion of multiple correlated variables in the model will make the individual regression coefficients to be estimated unstably, i.e. with big variance. If we only use a small subset of variables, then the models are likely to be biased. Often we would like to achieve a balance between the variance and the bias [2]. This project analyzes the Motor Trend Car Road Tests (mtcars) dataset, a collection of cars in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome), We are trying to answer the following two questions:
1. Is an automatic or manual transmission better for MPG?
2. Quantify the MPG difference between automatic and manual transmissions.
Two models are selected to demonstrate the importance of model selection, one is automatic/manual as the only predictor, and the other is the minimal adequte model.

Exploratory Data Analysis

The data look clean and there are no missing values to impute (Appendix A). However we need to transform some of the variables to factor variables (cyl, vs, am, gear, carb). The table shows the first 5 observations of the mtcars dataset.

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
1	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
2	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
3	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
4	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
5	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2

Regression Analysis

If we naively took "am" (0 automatic, 1 manual) as the predictor and "mpg" as the response, we may mistakenly conclude that, on average, the manual transmission cars have 7.25 higher MPGs than automatic transmission cars as shown in the table below. However if you look at the "Residuals Fitted (1)" plot in Appendix B, there is a clear pattern. Also, the R^2=0.36 is small, in other words, the model considering single predictor "am" only explains 36% of the variability of the response data around its mean. There are confounders that we need to include into the regression model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.147	1.125	15.247	0
am1	7.245	1.764	4.106	0

Model Selection

We begin with throwing all the predictors into the regression model and performs stepwise model selection by AIC (Appendix C). The final model is using "cyl", "hp", "wt" and "am" as the predictors. The coefficients are shown in the following table. The "Residuals vs Fitted (2)" plot in Appendix B does not show any obvious patterns.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.708	2.605	12.94	0
cyl6	−3.031	1.407	−2.154	0.041
cyl8	−2.164	2.284	−0.947	0.352
hp	−0.032	0.014	−2.345	0.027
wt	−2.497	0.886	−2.819	0.009
am1	1.809	1.396	1.296	0.206

The slope of "am" is 1.809, which means, on average, holding other variables constant the manual transmission cars have 1.809 higher MPGs than automatic transmission cars. However the "p-value" (0.206) is large, so we failed to reject the null-hypotheis that the manual transmission cars have the same MPGs as that of automatic transmission cars. The 95% confidence interval is [-1.061, 4.679] also includes 0.

Summary

After analyzing the "mtcars" dataset, we conclude:
1. There is no significant diffrence on MPG by automatic or manual transmission,
2. The MPG difference between automatic and manual transmissions is [-1.061, 4.679] at 95% C.L.

Reference

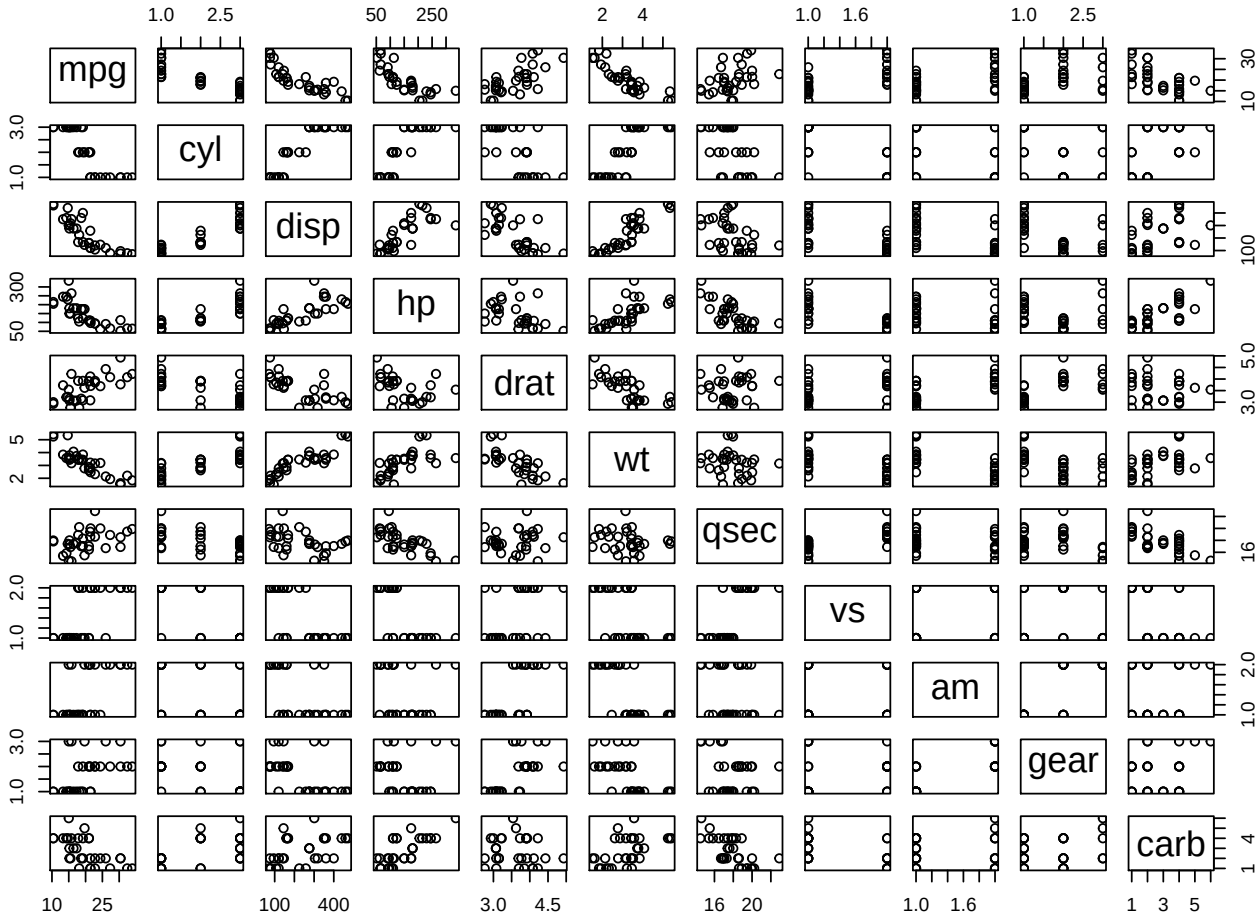
[1] Jianqing Fan and Jinchi Lv, Stat Sin. 2010 Jan; 20(1): 101–148.
[2] http://www.ams.sunysb.edu/~xing/AMS578/Documents/Chp9.pdf

Apendeix

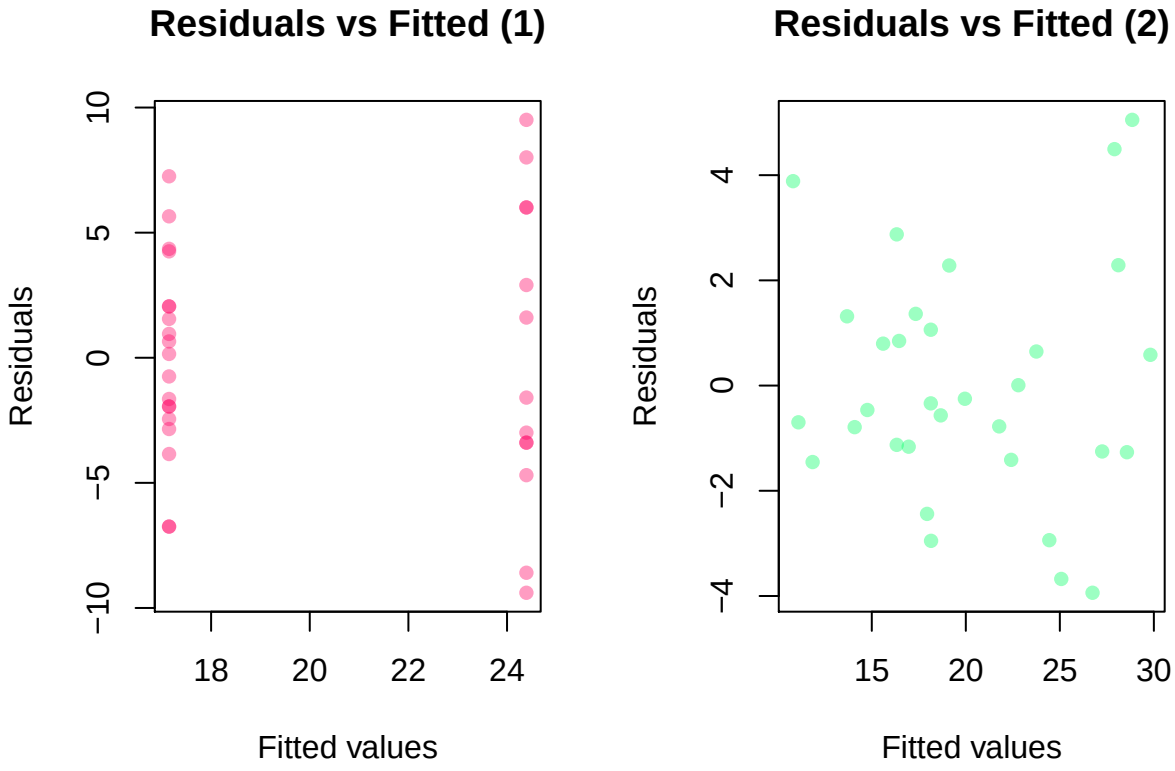
Apendeix A

```
## '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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

```
##       mpg        cyl         disp             hp             drat      
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760  
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080  
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695  
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597  
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920  
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930  
##        wt             qsec       vs     am     gear   carb      
##  Min.   :1.513   Min.   :14.50   0:18   0:19   3:15   1: 7  
##  1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   4:12   2:10  
##  Median :3.325   Median :17.71                 5: 5   3: 3  
##  Mean   :3.217   Mean   :17.85                        4:10  
##  3rd Qu.:3.610   3rd Qu.:18.90                        6: 1  
##  Max.   :5.424   Max.   :22.90                        8: 1
```


mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb

Apendix B



Appendix C

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##
## Final Model:
## mpg ~ cyl + hp + wt + am
##
##
##     Step Df   Deviance Resid. Df Resid. Dev      AIC
## 1                             15   120.4027 76.40339
## 2 - carb  5 13.5988573        20   134.0015 69.82769
## 3 - gear  2  5.0215145        22   139.0230 67.00492
## 4 - drat  1  0.9672159        23   139.9903 65.22678
## 5 - disp  1  1.2473996        24   141.2377 63.51066
## 6 - qsec  1  2.4420033        25   143.6797 62.05921
## 7   - vs  1  7.3459298        26   151.0256 61.65483
```