DDS_RegMod_w4 submission

Jie Yang

Executive Summary

Questoin 1: Is an automatic or manual transmission better for MPG?

Exploratory data

```
str(mtcars)
```

In total, this dataset has 32 observations, with 11 variables.

Key parameters needed in this task are mpg (mile per gallon), am (transmission): 0 strands for automatic and 1 stands for mannual. We need to do linear regression on this. Plots are available in the Appendix.

Model Selections:

Since the mpg is continous, and am is either 0 or 1, we can use linear regression `lm(mpg~ ., data=mtcars)`. However, one need to find the regressor that can explain the mpg. With the `step()` function (or `drop1()`,mannual examination on p value for F-test), one can find that the transmission (am), weight (wt), 1/4 mile time(qsec) are used for linear regression, where each coeffient is different from zero at confidence level of 95%. Ovearll, this model can explain 83% of the variance and the resiual plots shows it's a suffient linear regression model. Details are avaialbe in Appendix, model selection. But essentically, one get

$$mpg_i = \beta_0 + \beta_1 qsec_i + \beta_2 wt_i + \beta_3 am_i + \epsilon_i$$

where

$$\beta_0 = 9.62, \beta_1 = 1.23, \beta_2 = -3.92, \beta_3 = 2.94$$

As the summary above shows, the coeffecients are

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

The simple one variable linear regression model `lm(mpg~am)` only explains 36% of the variance from mpg.

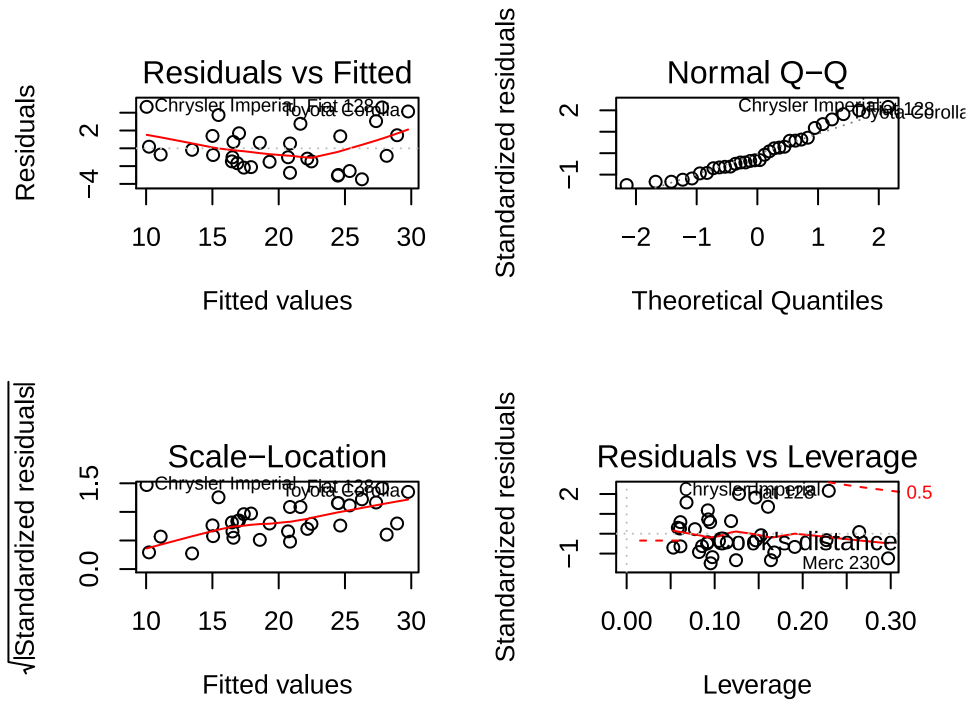
$$mpg_i = \beta_0 + \beta_1 am_i + \epsilon_i$$

More details on model selection is avaiable at appendix.

Residual Anallysis and Plots

For linear regression, we need residual plots, to test if there's residual is depedence on x or y in the model.

```
par(mfrow=c(2,2)) # alterntive could be plot(predict(ffit),resid(ffit),pch='*');abline(h=mean(resid(ffi
plot(ffit)
```



From the plot, we see 1. Residual vs fitted doesn't show consistent pattern (no heteroscedasticity) 2. Normal Q-Q plots shows residual are normally distributed as the dots lie on the line largely. 3. Scale location plots suggest constant variance as dots are randomly located. 4. Residual leverage suggests no outlier as they are within 0.5 band.

Interpretations of coeffcients

Now, we get the coeffients after model selections. Each p value is <0.05, they're significantly different from zero with confidence level of 95%. It explains 83% (adjusted R squared) of the total variance.

```
ffit<-lm(mpg ~ qsec+wt+am, data=mtcars)
summary(ffit)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  9.617781  6.9595930  1.381946 1.779152e-01
## qsec         1.225886  0.2886696  4.246676 2.161737e-04
## wt          -3.916504  0.7112016 -5.506882 6.952711e-06
## am           2.935837  1.4109045  2.080819 4.671551e-02
```

Question 2

"Quantify the MPG difference between automatic and manual transmissions"

Now, from Question 1, we decided the model(ffit)

$$mpg_i = \beta_0 + \beta_1 qsec_i + \beta_2 wt_i + \beta_3 am_i + \epsilon_i$$

where

$$\beta_0 = 9.62, \beta_1 = 1.23, \beta_2 = -3.92, \beta_3 = 2.94$$

The difference introduced by automic and manual transmission is from 2,94, so manual will increaes the mpg by 2.94 mile per gallon compared with automatic. The confidence interval is 0.05~5.83.

```r
confint(ffit, "am")
```

```
##         2.5 %   97.5 %
## am 0.04573031 5.825944
```

Appendix:

1. Hypotheis test

Based on the exploratory information from above, we'll generat the hypothesis that

H0: automatic is equal to manual Ha: automatic is less than the manual

To do a test

```r
t.test(mtcars[mtcars$am==0,]$mpg,mtcars[mtcars$am==1,]$mpg, alternative = "less", paired = F, var.equal
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars[mtcars$am == 0, ]$mpg and mtcars[mtcars$am == 1, ]$mpg
## t = -3.7671, df = 18.332, p-value = 0.0006868
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -3.913256
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

Given the pvalue =0.0006868, we reject the H0 hypothesis, so we conclude that automatic is less than manual (with less mpg). The confidence level is 95%, and the confidence interval for the difference (auto-manual in mpg) is

```
95 percent confidence interval:
      -Inf -3.913256
```

2. More details on the model selections:

```r
#fit all the parameters in the lm
model_full<-lm(mpg ~ . , data=mtcars)
summary(model_full)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
```

```
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

Although the total p-value is low, each coeffcient p value is not less than 0.05. So, none of them are signficant in the confidence range we use in this report 95%. We resort to the `step()`function to find the ones sigifnicifant. It will list them in ascending order. Since the smaller AIC value is more likely to resemble the TRUTH model

```
step_fit<-step(model_full)
summary(step_fit)
```

Alternative methods are `drop1()`,`add1()`(using F test for multiple variable linear regressions)

```
drop1(model_full, test="F")
#above find cyl to drop, as the p value is 0.91609
drop1(update(model_full, ~ . -cyl), test = "F")
#above find disp to drop, as the p value is 0.45381
drop1(update(model_full, ~ . -cyl -disp), test = "F")
#above find vs to drop, as the p value is 0.96332
drop1(update(model_full, ~ . -cyl -disp -vs), test = "F")
#copying above, we'll continue to drop a few parm: drat, gear, hp, carb
drop1(update(model_full, ~ . -cyl -disp -vs -drat -gear -hp -carb), test = "F")
#This is finally agreeing with the step AIC based approach, we finally keep am, qsec,wt
```

Automatic F test based approach:

```
library(rms)
ols.full <- ols(mpg ~ cyl+disp+hp+drat+wt+qsec+vs+am+gear+carb, data=mtcars)
fastbw(ols.full, rule = "p", sls = 0.05)
```

Above approach only leaves wt, qsec (ref.7). But for this study, we can keep am, as the p value is <0.05 (confi.level 95%). Anova also suggest am helps to explain the mpg.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + qsec
## Model 2: mpg ~ wt + qsec + am
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     29 195.46
## 2     28 169.29  1    26.178 4.3298 0.04672 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3. Further refinements of the model (Optional): As the pair plot suggest, we can see there could be interactions between weights and am, qsec and am.

`t.test(mtcars[mtcars$am==0,]$wt,mtcars[mtcars$am==1,]$wt, alternative = "greater", paired = F, var.equal = F, conf.level = 0.95)` suggest p-value = 3.136e-06 to reject H0 .

But regarding qsec, `t.test(mtcars[mtcars$am==0,]$qsec,mtcars[mtcars$am==1,]$qsec, alternative = "two.sided", paired = F, var.equal = F, conf.level = 0.95)`, since `p-value = 0.2093` we cannot reject the *H0* that mean are identical on qsec.

Hence, we can further include the interactions term `wt:am`. The model indeed is better from the pvalue in nova (0.001809<<0.05), and adjusted R2 (ffit is 0.83, ffit is 0.88, fit is only 0.36)

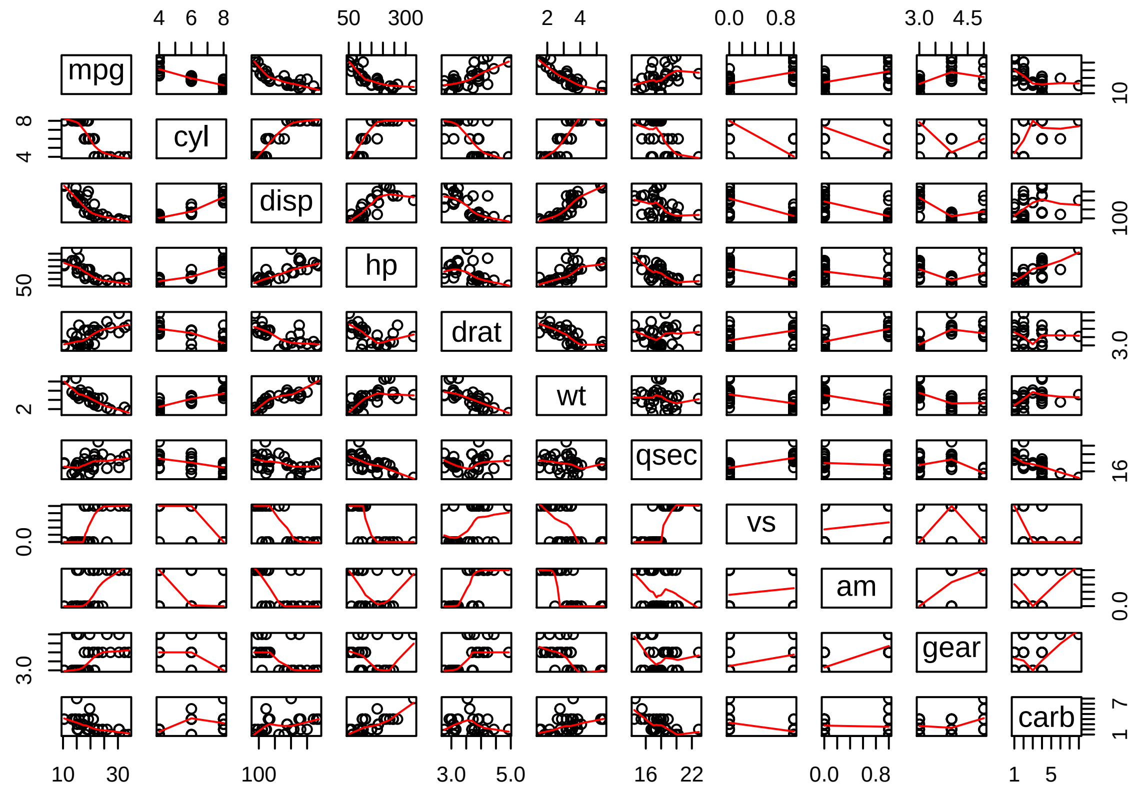
```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am + am:wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am            14.079      3.435   4.099 0.000341 ***
## wt:am         -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

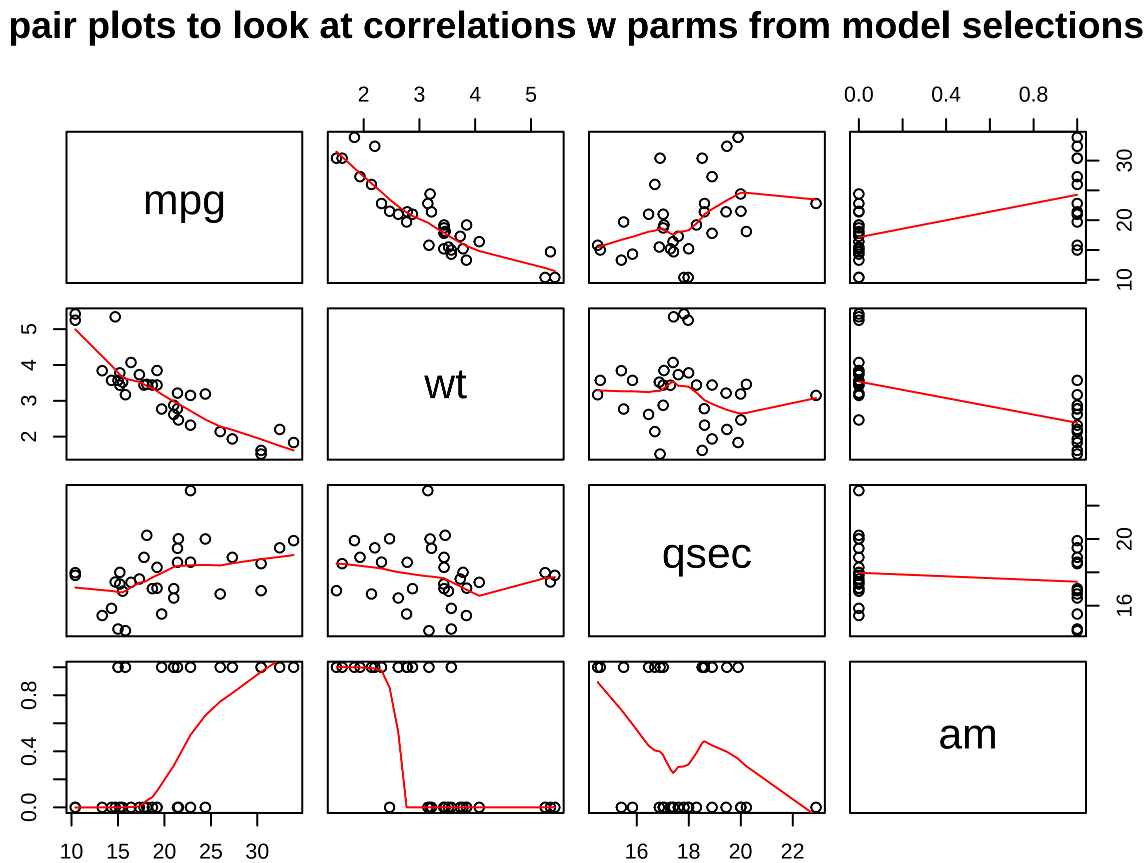
```
## Analysis of Variance Table
##
## Model 1: mpg ~ qsec + wt + am
## Model 2: mpg ~ wt + qsec + am + am:wt
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     28 169.29
## 2     27 117.28  1     52.01 11.974 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. Pair plot to show the correlations

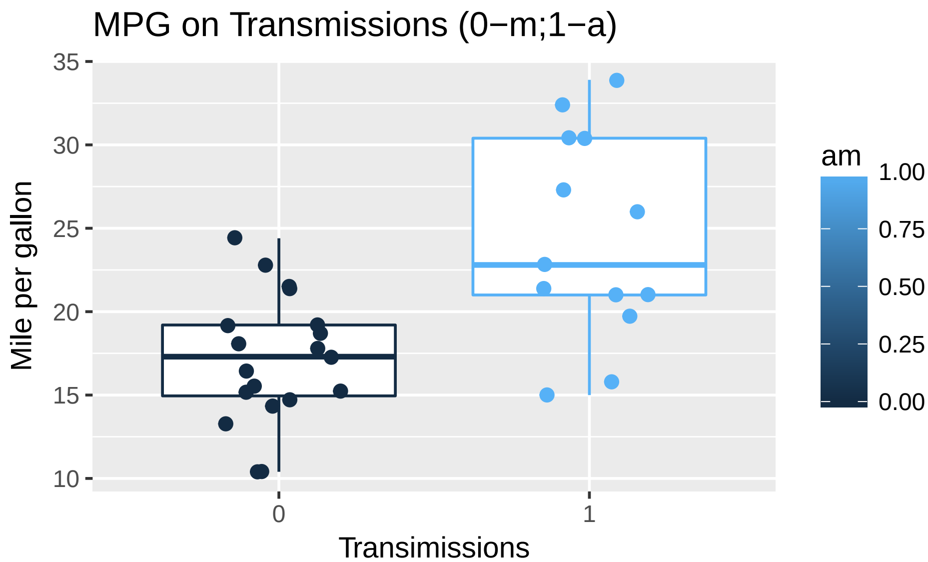
```
## Warning in par(frow = c(2, 1)): "frow" is not a graphical parameter
```

Overview of correlations





5. Exploratory Plots for the median for transmissions.



Reference works:

1. https://github.com/alex23lemm/Regression-Models-Project/blob/master/mtcars_analysis.pdf
2. https://github.com/codebender/regression-models-course-project/blob/master/Motor%20Trend%20MPG%20Data%20Analysis.pdf
3. https://github.com/fcampelo/RM-course-project
4. https://github.com/Xiaodan/Coursera-Regression-Models/blob/master/motor_trend_project/report.pdf
5. https://stats.stackexchange.com/questions/214682/stepwise-regression-in-r-how-does-it-work
6. https://stat.ethz.ch/R-manual/R-devel/library/stats/html/step.html
7. http://rstudio-pubs-static.s3.amazonaws.com/2899_a9129debf6bd47d2a0501de9c0dc583d.html
8. https://stats.stackexchange.com/questions/172782/how-to-use-r-anova-results-to-select-best-model