Assignment: mtcars dataset analysis

Dung Nguyen

July 15, 2018

Execute summary

The automobile magazine *Motor Trend* wants to have some insights about what affects the fuel efficiency so we are going to explore available variables that might influence miles per gallon (*mpg*). In this particular example, we are going to explore 2 questions. First, we would like to know whether an automatic or manual transmission is better for MPG. We also want to quantify the MPG difference between automatic and manual transmissions.

Data

We use the *mtcars* dataset for the analysis. Here is the quick summary of the data:

```
data(mtcars)
str(mtcars)
```

```
## '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 ...
```

We first explore and look for any clear relationships between variables using a pair plot matrix (Figure 1).

According to the Figure 1, Weight (*wt*) appears to be an important variable when it seems to have a pretty clear linear relationship with our dependent variable *mpg*. It makes sense too that the weight of the vehicle affects its miles per gallon since the heavier the vehicle is, the more fuel it will cost to travel at the same distance, hence the reverse relationship with *mpg*. Therefore, this is definitely an important variable that should be considered.

Displacement (*disp*) and Horsepower (*hp*) demonstrate a good relationship with MPG as well, although 2 variables seem to have "curvy" relationships with *mpg*, and they themself appears to correlate with each other hence more careful analysis might need to explore whether these 2 variables should be used together to explain our dependent variable *mpg*. In theory, it is reasonable to think either of the variables (or both) might be very important since miles per gallon is definitely a variable that can be modelled based on the engine's capacity. It's worth noting that *disp* may have strong correlation with *wt* while *hp* has a relatively unclear relationship with *wt*.

Variable *cyl* contains 3 levels of 4, 6, or 8, which apear to have some significant difference of miles per gallon for each group on average, although these differences may have something to do with *wt* according to Figure 1. *cyl* might translate into weight, thus partly captured in *wt*.

Does transmission (*am*) matters?

From the plot Figure 2, it appears that manual cars ($am = 1$) may have higher *mpg*, or better at fuel efficiency in other words. Let's test the hypothesis that manual cars have higher miles per gallon compared to automatic cars and quantify this difference in terms of miles per gallon:

```
summary(lm(mpg~am, data = mtcars))
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

Linear regression suggests that manual cars ($am = 1$) have 7.245 miles per gallon higher than automatic cars do. The test's results suggests that manual cars have significantly greater *mpg* compared to automatic cars ($p < 0.001$). In general, compared to automatic cars, manual cars have higher *mpg* or better fuel efficiency. Moreover, tranmission type explains about 34% of the variance in fuel efficiency (*mpg*).

Covariates

Based on the exploratory plot Figure 1, it might benefit to transform a few variables in advance before doing some in-depth analysis.

Many previous analyses, using stepwise model selection (AIC metric), suggest that a combination of *qsec*, *wt*, *am* best predicts *mpg* with adjusted R-squared of approximately 84%. I believe that this analysis is good, yet it's strange somehow that it does not take engine's capacity into account at all, which was demonstrated to have somewhat nonlinear relationships with *mpg*.

According to Figure 1, it may makes sense to take a log of *mpg*, a positive variable. For the potential curviness, either a log transformation for these variable or a higher degree polynomial function (second) can be considered. Both transformations can be utilized, but imply different behaviors at extrapolation. I guess a log is more appropriate here based on the shape of the exploratory plot as well as the skewness first, but also it makes sense in terms of the law of diminishing return: more horsepower or displacement can mean more gallons are used at a time, yet the differential may get smaller. It's harder to imagine this relationship has an extrema where the relationship get reverse after the point of extrema, which is the case of polynomial function.

I create the pair plot matrix again, with log-transformed variables added in Figure 3. According to the plot, *log_mpg* has a pretty straight linear relationship with *wt* and *log_disp* and *log_hp*. Therefore, the transformation I make is the log(mpg), and the log(hp), log(disp). Note that there is a correlation between *log_hp* and *wt*. Let's see what variables stepwise regression will pick with newly transformed variables.

```
library(MASS)
mtcars$log_mpg <- log(mtcars$mpg);
mtcars$log_hp <- log(mtcars$hp)
mtcars$log_disp <- log(mtcars$disp)
step = stepAIC(lm(log_mpg~cyl+log_disp+log_hp+drat+wt+qsec+vs+am+gear+carb,
                  data = mtcars))
```

```
step[[13]]
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## log_mpg ~ cyl + log_disp + log_hp + drat + wt + qsec + vs + am +
##     gear + carb
##
## Final Model:
## log_mpg ~ log_hp + wt
##
##
##         Step Df     Deviance Resid. Df Resid. Dev       AIC
## 1                                   21  0.2738444 -130.3498
## 2       - am  1 3.346077e-05        22  0.2738778 -132.3459
## 3     - drat  1 1.596658e-04        23  0.2740375 -134.3272
## 4       - vs  1 3.770279e-04        24  0.2744145 -136.2832
## 5     - qsec  1 3.708035e-03        25  0.2781226 -137.8537
## 6      - cyl  1 4.613621e-03        26  0.2827362 -139.3273
## 7 - log_disp  1 6.950471e-03        27  0.2896867 -140.5501
## 8     - carb  1 1.258158e-02        28  0.3022682 -141.1896
## 9     - gear  1 1.327271e-02        29  0.3155409 -141.8145
```

Stepwise regression suggests a rather simple model with only *wt* and *log_hp*. This result closely fit our theory and simple enough for a useful model. Let's run the regression with just these 2 variables and with 2 variables plus transmission (*am*)

```
fit <- lm(log_mpg~wt+log_hp, data = mtcars)
summary(fit)
```

```
##
## Call:
## lm(formula = log_mpg ~ wt + log_hp, data = mtcars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16296 -0.07799 -0.02210  0.06837  0.25985
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.83167    0.22198  21.766  < 2e-16 ***
## wt          -0.17942    0.02742  -6.543 3.63e-07 ***
## log_hp      -0.26566    0.05646  -4.706 5.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1043 on 29 degrees of freedom
```

```
## Multiple R-squared:  0.8852, Adjusted R-squared:  0.8773
## F-statistic: 111.8 on 2 and 29 DF,  p-value: 2.338e-14
```

The regression shows that both variables belongs to the model with coefficents significantly different from 0. Together, these variables explain about 88% of *log_mpg*. Although these variables are correlated with each other, their effects computed by least square are still strongly present. Let's check their variance inflation factor (VIF):

```
library(car)
vif(fit)
```

```
##     wt log_hp
## 2.0509 2.0509
```

Both VIFs are very low, suggesting the collinearity problem is not strong and does not need further attention.

There is no clear pattern in the residual plot Figure 4, and the residual looks approximately normal. However, there exists a point with very high Cook's distance compared to all other points: Chrysler Imperial car. It might be reasonable to try excluding this point to examine the robustness of our model::

```
summary(lm(log_mpg~wt+log_hp,
           data = mtcars[-which.max(cooks.distance(fit)),]))
```

```
##
## Call:
## lm(formula = log_mpg ~ wt + log_hp, data = mtcars[-which.max(cooks.distance(fit)),
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14877 -0.07235 -0.02257  0.07181  0.20863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.83070    0.19363  24.948  < 2e-16 ***
## wt          -0.20855    0.02561  -8.143 7.27e-09 ***
## log_hp      -0.24832    0.04955  -5.012 2.68e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09099 on 28 degrees of freedom
## Multiple R-squared:  0.9133, Adjusted R-squared:  0.9071
## F-statistic: 147.5 on 2 and 28 DF,  p-value: 1.356e-15
```

The result , in fact, a better fit, with 2 variables accounting for about 91% of the variability of *log_mpg*. According to the model, for 1000lbs increase in weight, fuel efficiency miles per gallon are expected to go down by 21%, holding horsepower constant. Similarly, for 1% increase in horsepower, miles per gallon are expected to go down 25%, keeping weight constant.

2 types of transmission (am) were found to be significantly different in terms of fuel effiency *mpg*, yet when added as a third variable to our model of 2 important variables, am became not significant. The coeffiencts of wt changes a little but not much, suggesting that the effect of transmission am is mostly captured in the other 2 variables. Figure 5 shows that automatic cars ($am = 0$) tends to weigh heavier than manual cars.

```
summary(lm(log_mpg~wt+log_hp+am,
           data = mtcars[-which.max(cooks.distance(fit)),]))
```

```
##
## Call:
```

```
## lm(formula = log_mpg ~ wt + log_hp + am, data = mtcars[-which.max(cooks.distance(fit)),
##     ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14977 -0.07104 -0.02118  0.07099  0.20811
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.831701   0.197837  24.423  < 2e-16 ***
## wt          -0.210115   0.036436  -5.767 3.92e-06 ***
## log_hp      -0.247253   0.053377  -4.632 8.18e-05 ***
## am          -0.003059   0.049819  -0.061    0.951
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09265 on 27 degrees of freedom
## Multiple R-squared:  0.9133, Adjusted R-squared:  0.9037
## F-statistic: 94.82 on 3 and 27 DF,  p-value: 1.877e-14
```

Appendix

```
library(PerformanceAnalytics)
chart.Correlation(mtcars, histogram=TRUE, pch=19)
```

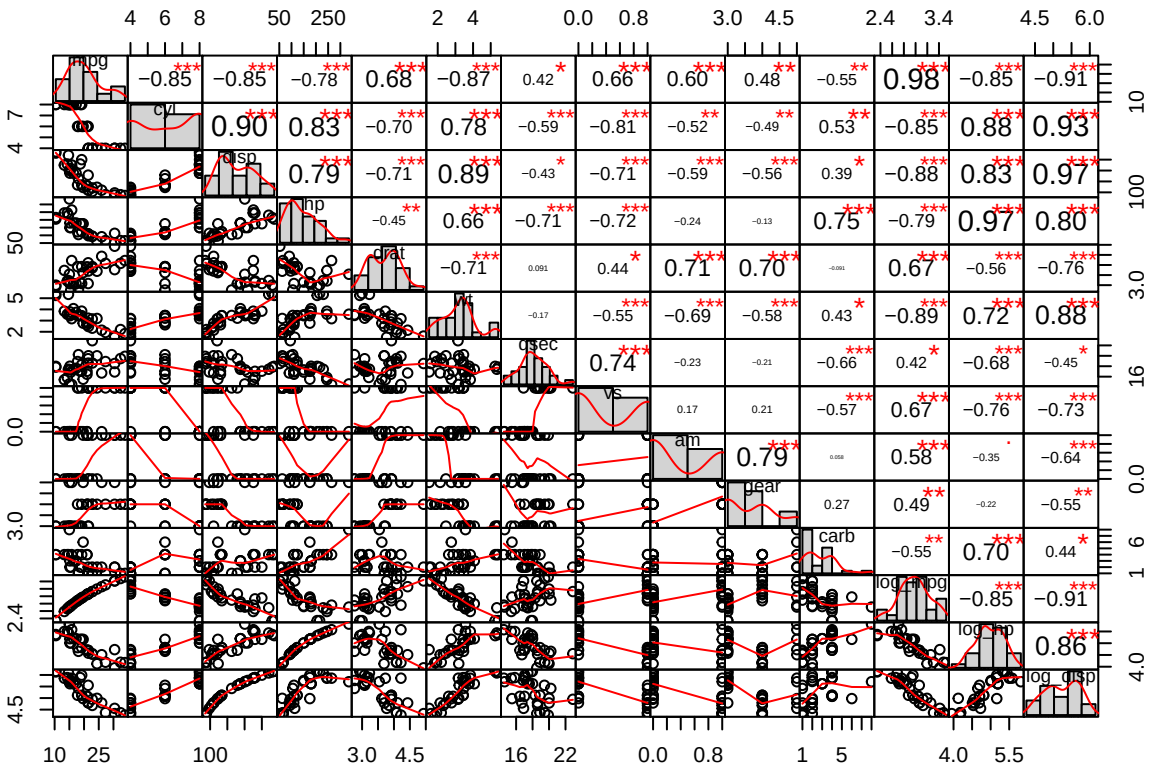

Figure 1: Scatterplot matrix

```
library(ggplot2)
ggplot(data = mtcars, aes(x=factor(am), y=mpg)) +
```

```
    geom_boxplot(fill = "lightblue", color = "black") +
    geom_jitter(position=position_jitter(width=.1, height=0)) +
    labs(x = "am")
```

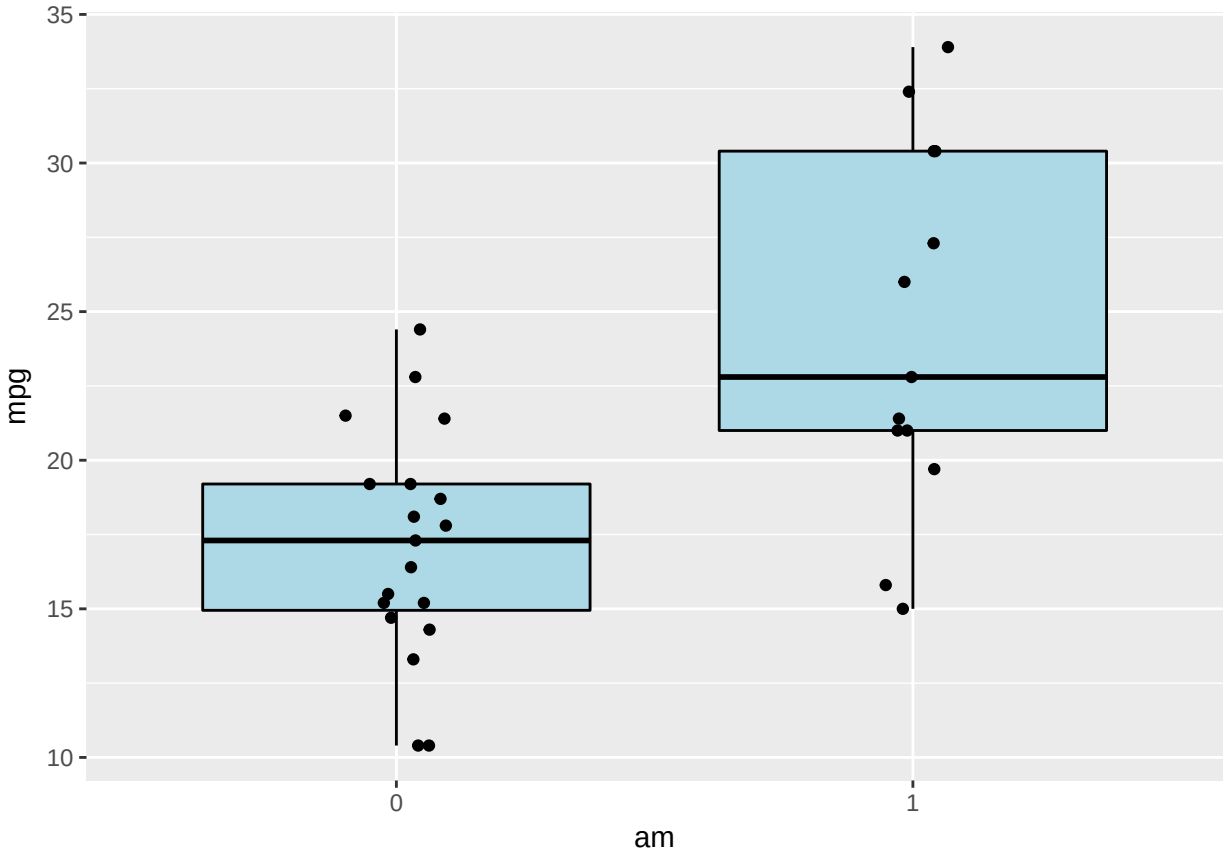

Figure 2: *mpg* difference between transmission types

```
chart.Correlation(mtcars, histogram=TRUE, pch=19)
```

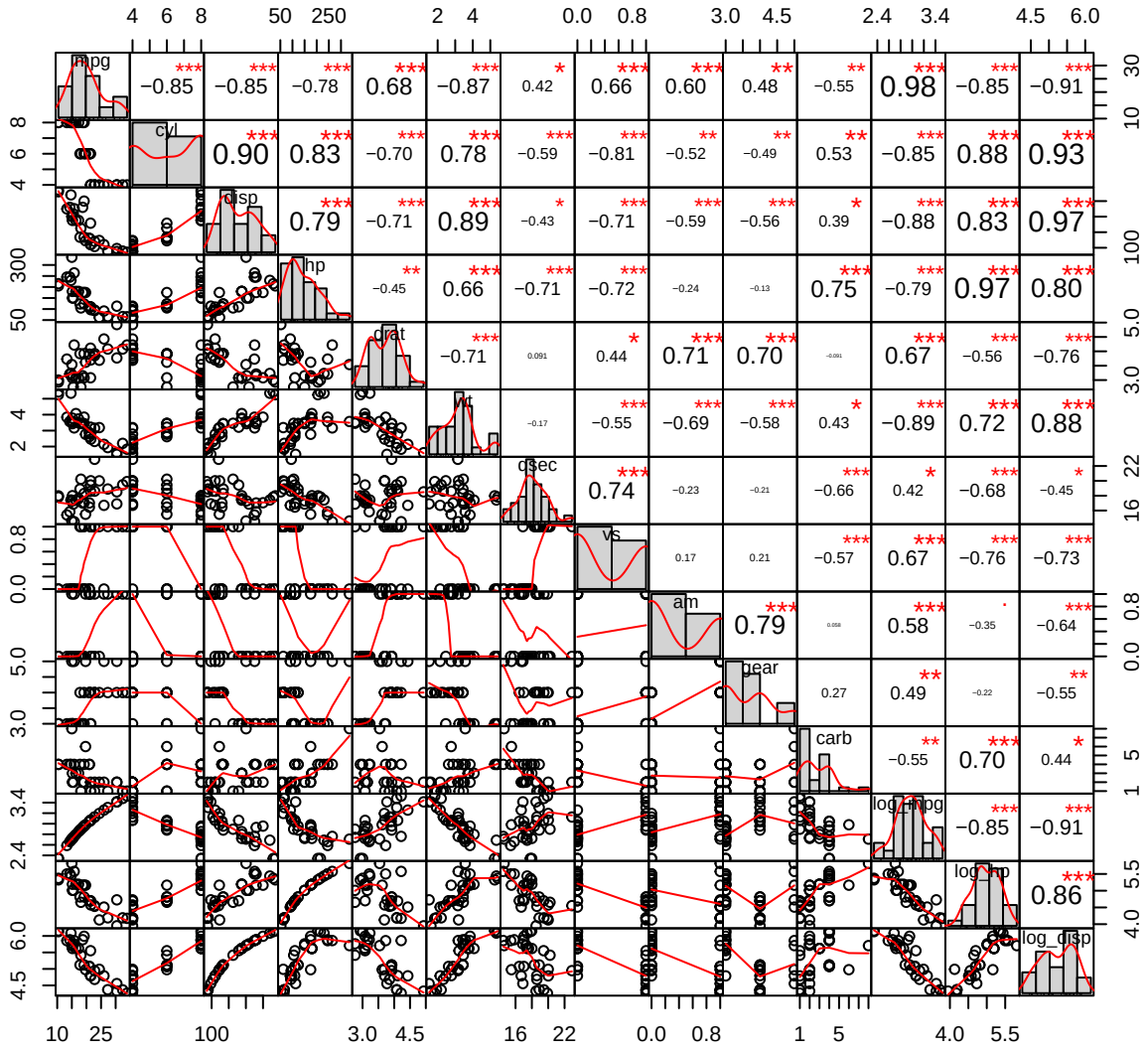

Figure 3: Scatterplot matrix with log_transformed variables added

```
par(mfrow = c(2,2))
plot(fit)
```

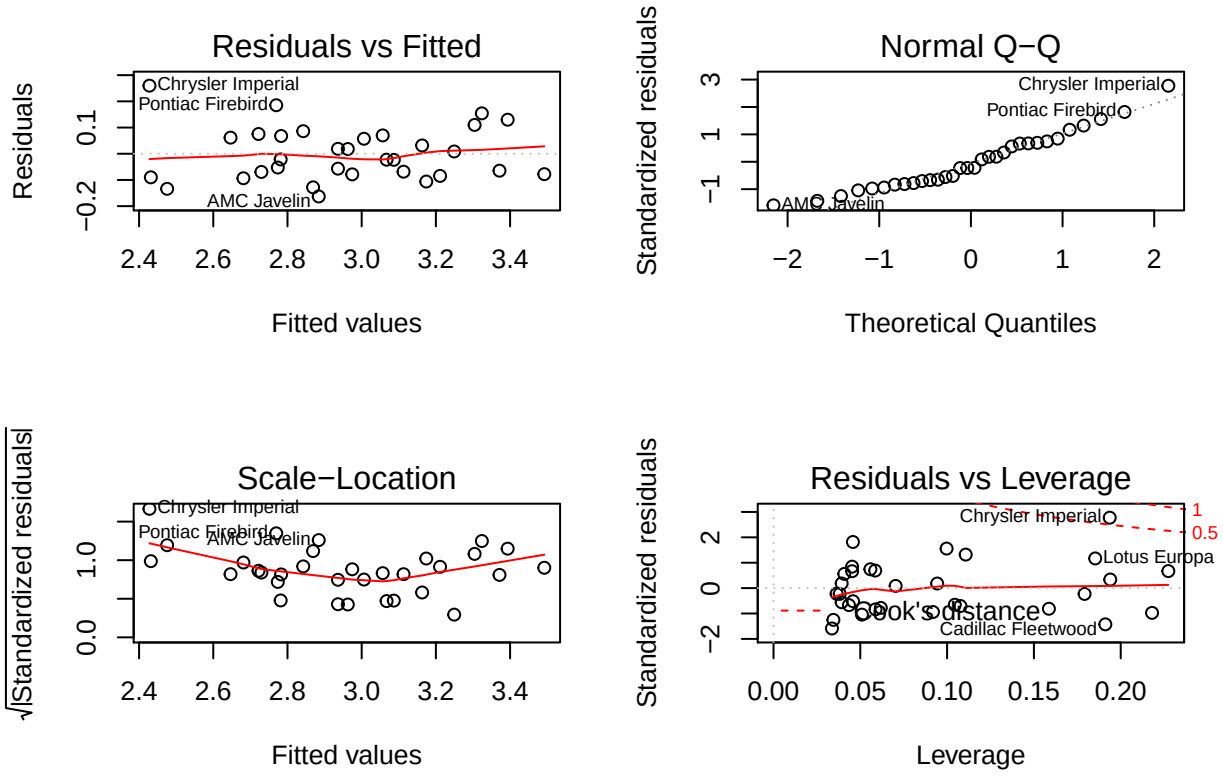

Figure 4: Residual and diagnostic plots

```
library(gridExtra)
plot1 <- ggplot(data = mtcars, aes(x=wt, y=log_mpg)) +
      geom_point(aes(color = factor(am))) +
      labs(x = "wt")
plot2 <- ggplot(data = mtcars, aes(x=log_hp, y=log_mpg)) +
      geom_point(aes(color = factor(am))) +
      labs(x = "log_hp")
grid.arrange(plot1, plot2, ncol = 2)
```

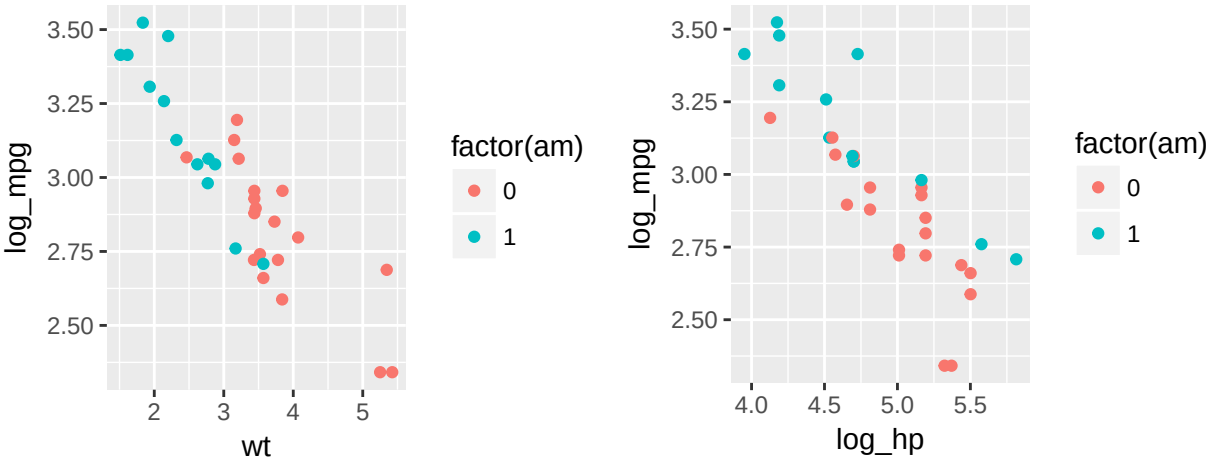

Figure 5: The effect of *am* is captured in *wt*, hence its effect is not significant when *wt* is present in the regression.