Assignment 1

1.

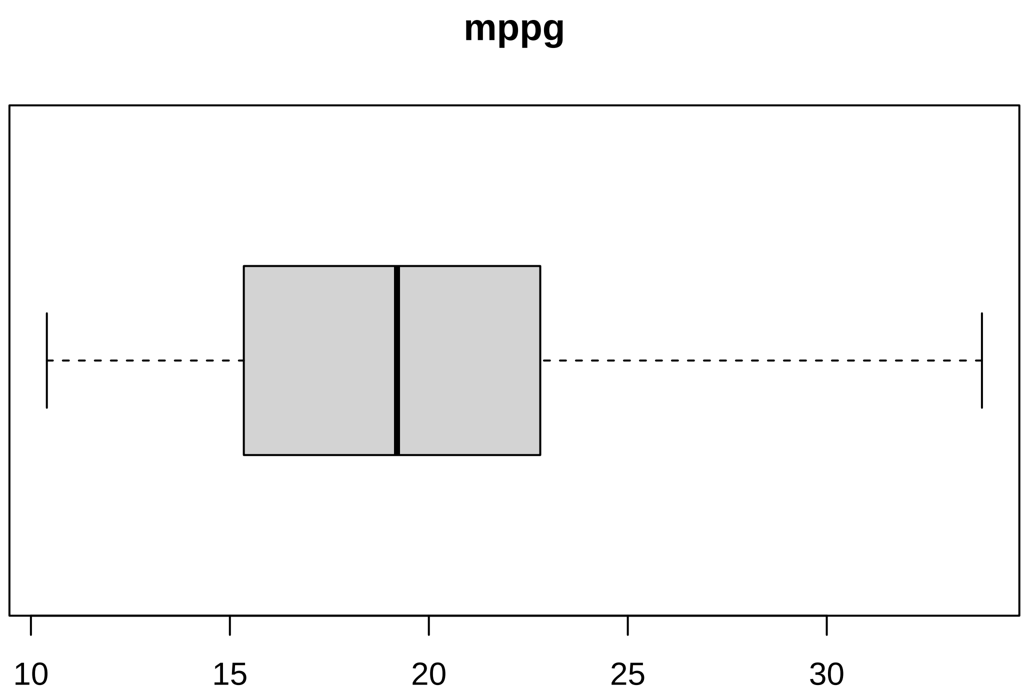
Our data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973-1974 models). The data has 32 observations on the following 11 variables:

mpg: Miles/(US) gallon
cyl: Number of cylinders
disp: Displacement (cubic inches)
hp: Gross horsepower
drat: Rear axle ratio
wt: Weight (1000 lbs)
qsec: 1/4 mile time
vs: Engine (0=V-shaped, 1=straight)
am: Transmission (0=automatic, 1=manual)
gear: Number of forward gears
carb: Number of carburetors

Our goal is to create a linear model with **mpg** as the response variable.
We start with some basic exploratory data analysis. The following is a basic summary of the data set.

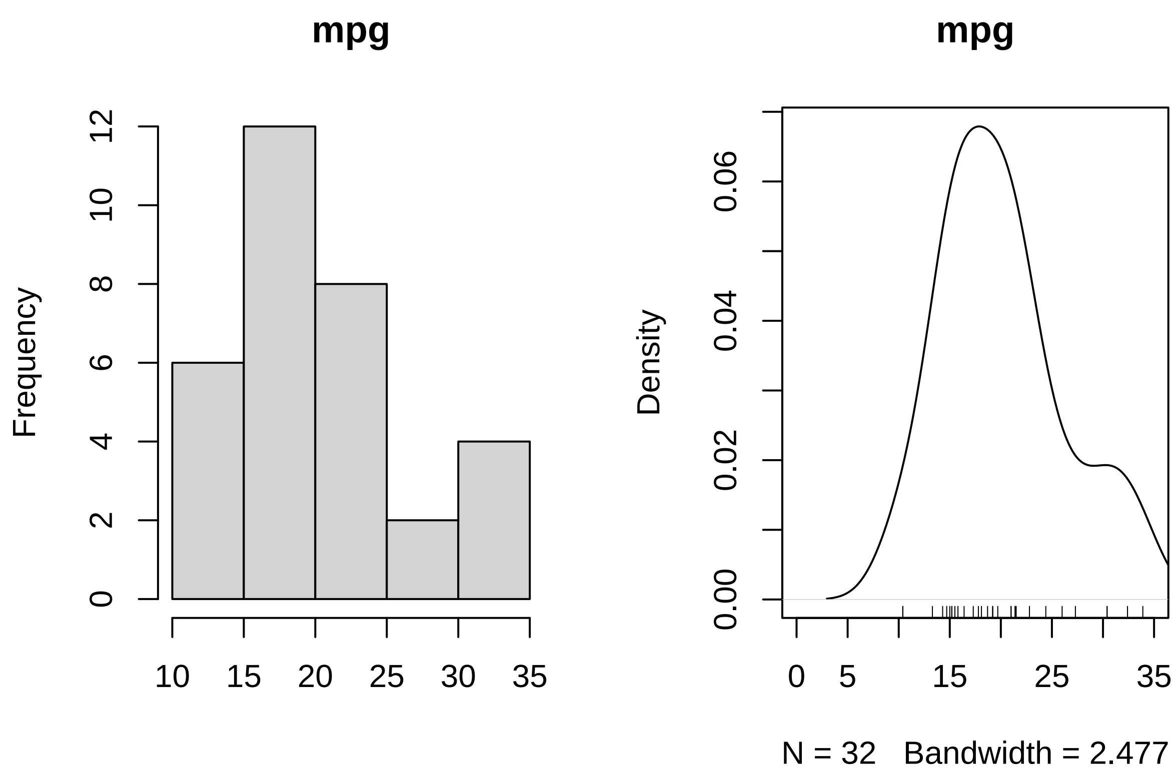
```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec        vs     am          gear
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   0:18   0:19   Min.   :3.000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1:14   1:13   1st Qu.:3.000
##  Median :3.695   Median :3.325   Median :17.71                 Median :4.000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85                 Mean   :3.688
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90                 3rd Qu.:4.000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90                 Max.   :5.000
##       carb
##  Min.   :1.000
##  1st Qu.:2.000
##  Median :2.000
##  Mean   :2.812
##  3rd Qu.:4.000
##  Max.   :8.000
```

This is a boxplot of **mpg**:



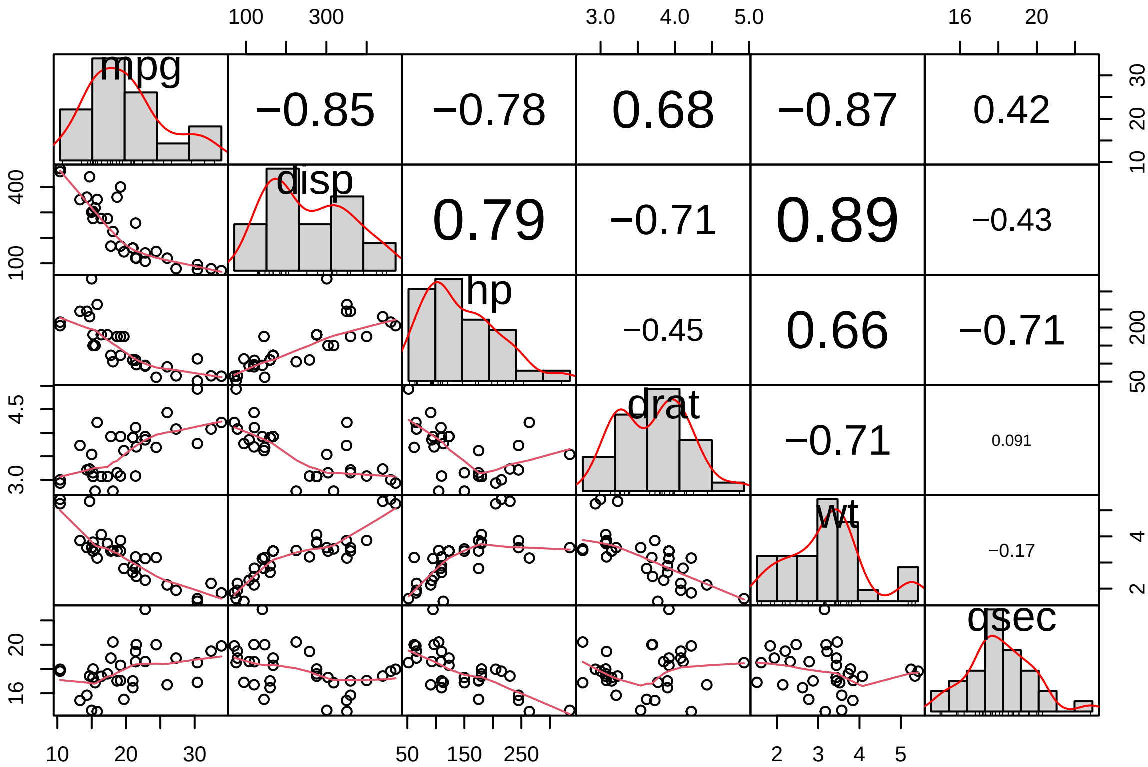
From this boxplot, we can see that most values of **mpg** are between 15 and 23, there is slight positive skew, and there are no major outliers. All the values are all of the same order of magnitude (all within 10-40).

These two figures are a histogram of mpg and an estimated density curve of mpg (with the values of mpg marked below the curve):



From these diagrams, we can see that **mpg** is unimodal, not symmetric, and slightly positively skewed. **mpg**'s skewness is in fact 0.610655, which indicates moderate positive skewness (the measure of skewness used here is the sample version of $E(\frac{X-\mu}{\sigma})^3$).

Next we look at some associations between **mpg** and other variables. This diagram shows pairwise scatterplots of **mpg** and other continuous variables.

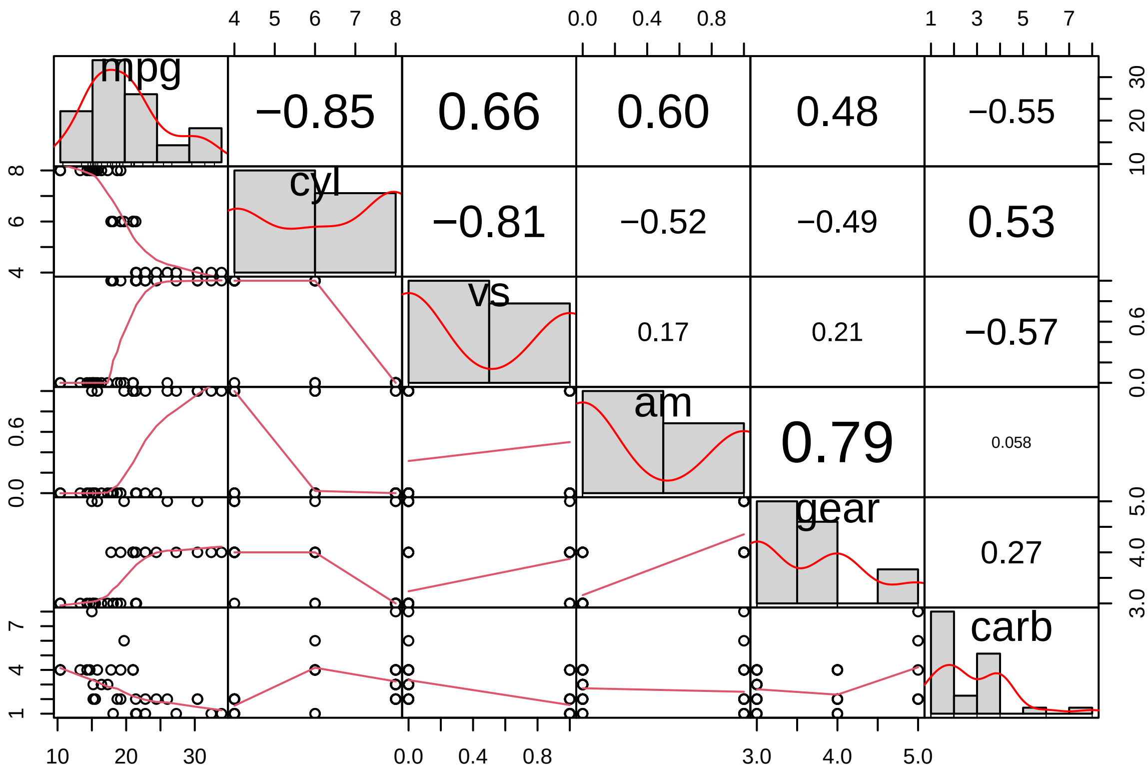


This diagram shows that the data is replete with (close to) linear associations. We are mainly interested in **mpg**, so we look for variables highly correlated with **mpg**.

mpg is in fact correlated with all of disp, hp, drat, wt, and qsec, with the lowest correlation being 0.42 (in absolute value), with qsec (indicating moderate correlation). The biggest correlation (in absolute value) is with wt, -0.87. This is not surprising at all since heavier cars would need more petrol to travel one mile. We can see that **mpg** is also highly negatively correlated with disp and hp and positively correlated with drat. **mpg** and hp having -0.78 correlation is to be expected, since cars with higher horsepower would be more powerful but with worse fuel economy.

Given such high correlations, these variables should perform well in a linear model with **mpg** as a response variable. However, we also see that disp, hp, and wt are themselves highly correlated, therefore there is multi-collinearity among these variables, and this will have to be dealt with.

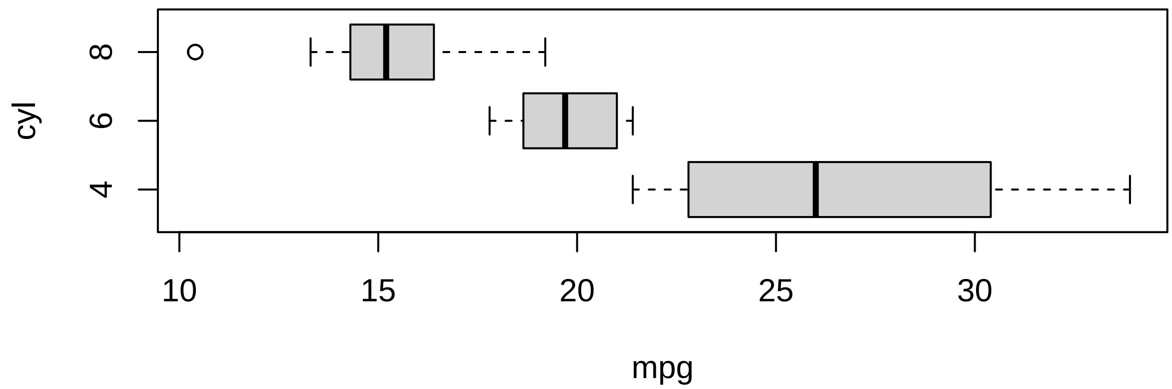
This diagram shows pairwise scatterplots of **mpg** and other discrete variables (with 0-1 coding for vs and am).

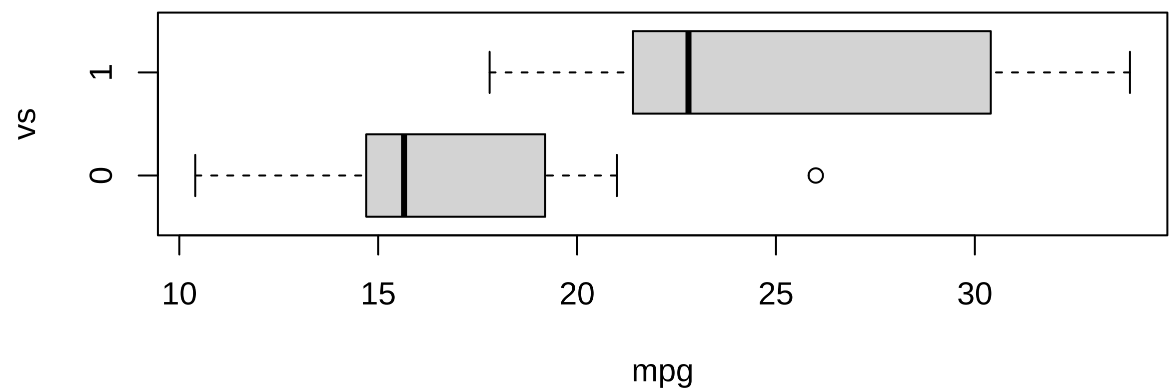


The biggest correlation in this diagram is between **mpg** and cyl; this is not surprising since an engine with more cylinders would be more powerful but consome more petrol.

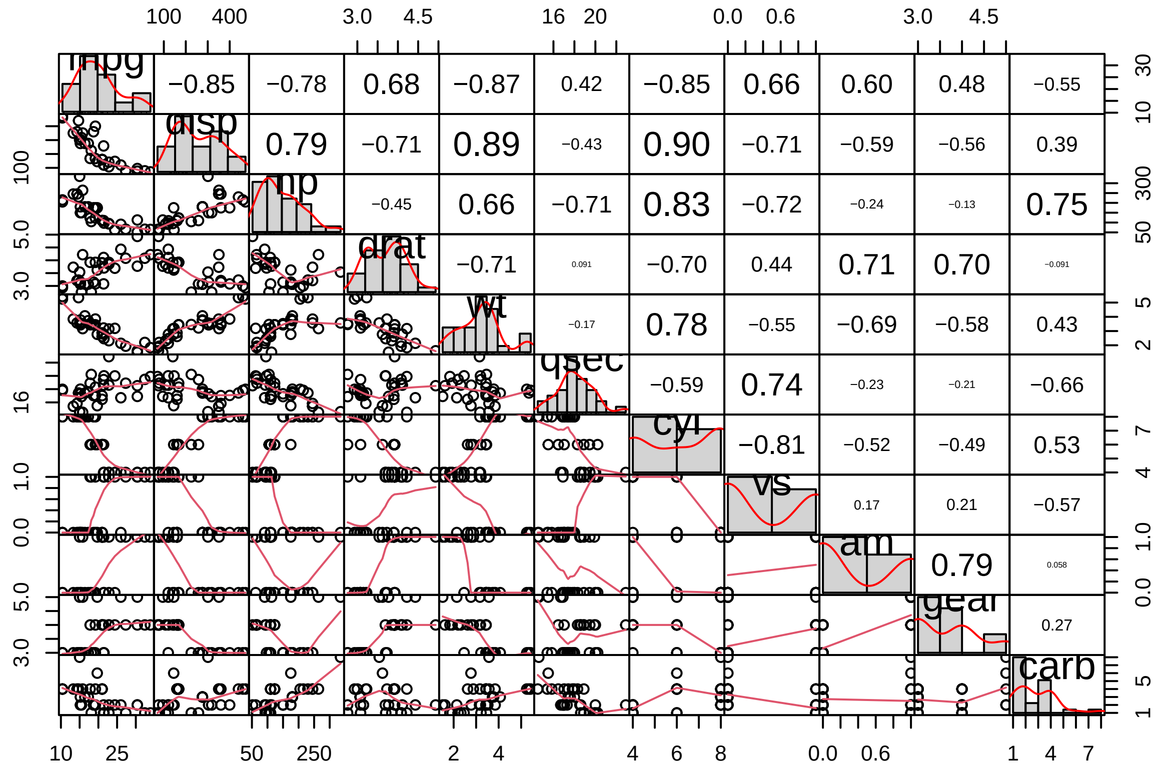
We also see that **mpg** is also substantially positively correlated with vs and am and moderately correlated with gear and carb.

cyl itself is also strongly negatively correlated with vs and moderately correlated with vs, am, gear, and carb. Given the high negative correlation between **mpg** and cyl, we can expect cyl to be a good predictor of **mpg**. We can also expect cyl to have multi-collinearity with vs, am, gear, and carb.





Here are box plots showing the relationship between **mpg** and the two discrete variables most correlated with it, cyl and vs. Clearly, cyl and vs would be good predictors of **mpg**, each of them could separate **mpg** into slightly overlapping intervals. However, cyl and vs themselves are also very highly correlated, so we probably will not need both variables in a linear model for **mpg**.



We include this diagram just to quickly check the correlation between the continuous and discrete variables. We can see that cyl and vs are both correlated with all the continuous variables (vs is less correlated with drat and wt, but still moderately correlated). Hence, there will be multi-collinearity if we don't perform any variable selection or transformation.

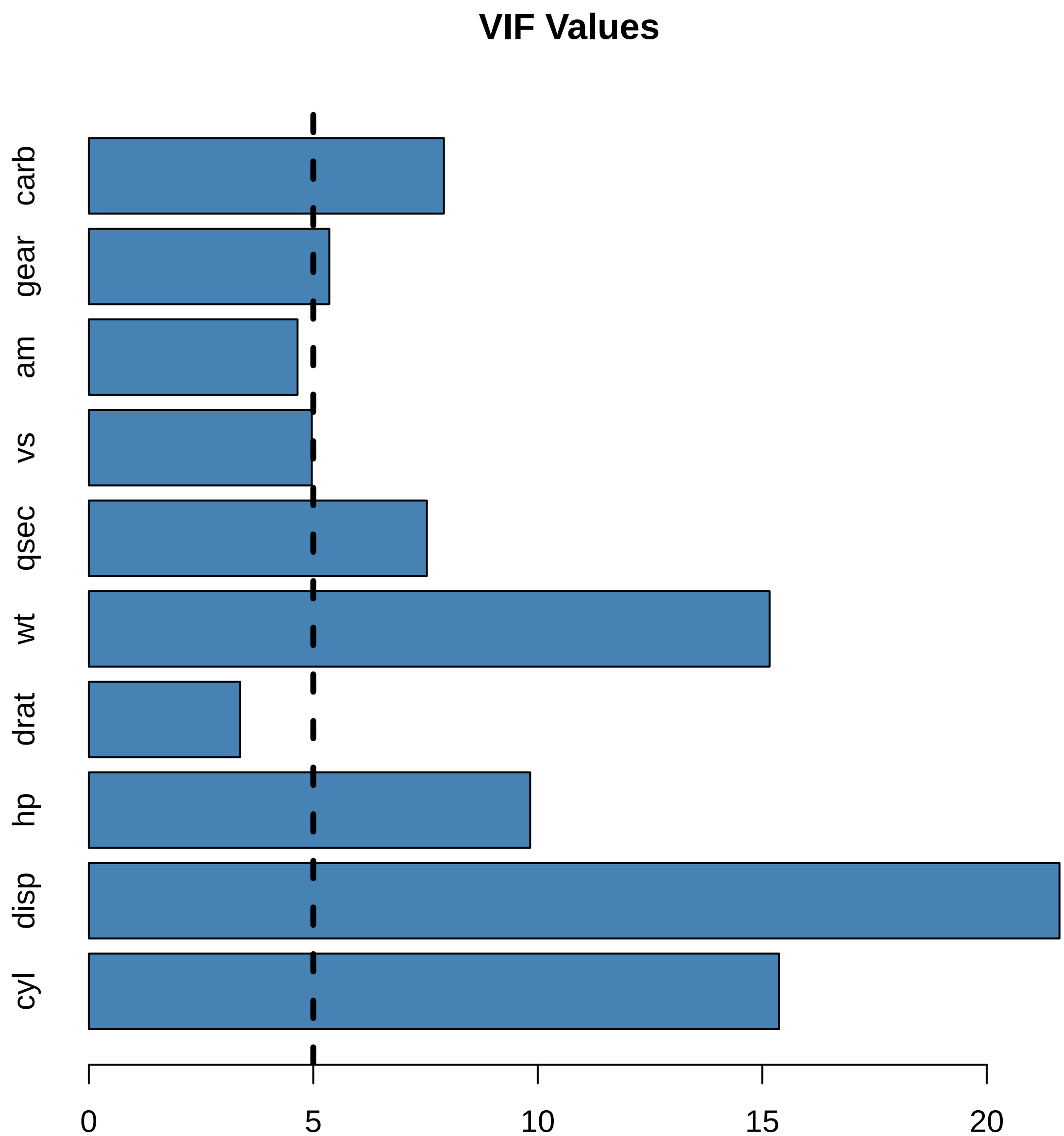
We fit the most basic linear model that is not trivial, the model including ALL variables:

$$Y = \beta_0 + \sum_{i=1}^{10} \beta_i \cdot X_i + \epsilon,$$

where Y is **mpg**, X_i's are the other variables, and $\epsilon \sim N(0, \sigma^2)$.

```
##
## Call:
## lm(formula = mpg ~ ., data = cars)
##
## 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
```

The p-value for the F-test is virtually zero, indicating that the explanatory variables do have a linear relationship with **mpg**. $R^2 = 0.869$ indicates that about 86.9 percent of the variance in **mpg** can be explained by the explanatory variables. However, none of the variables is significant. This is surely because of multi-collinearity, the influence of the explanatory variables mask each other, this is undesirable because we want a smaller model with only important variables. To get an idea of how much the model is suffering from multi-collinearity, we can look at the variance inflation factors:

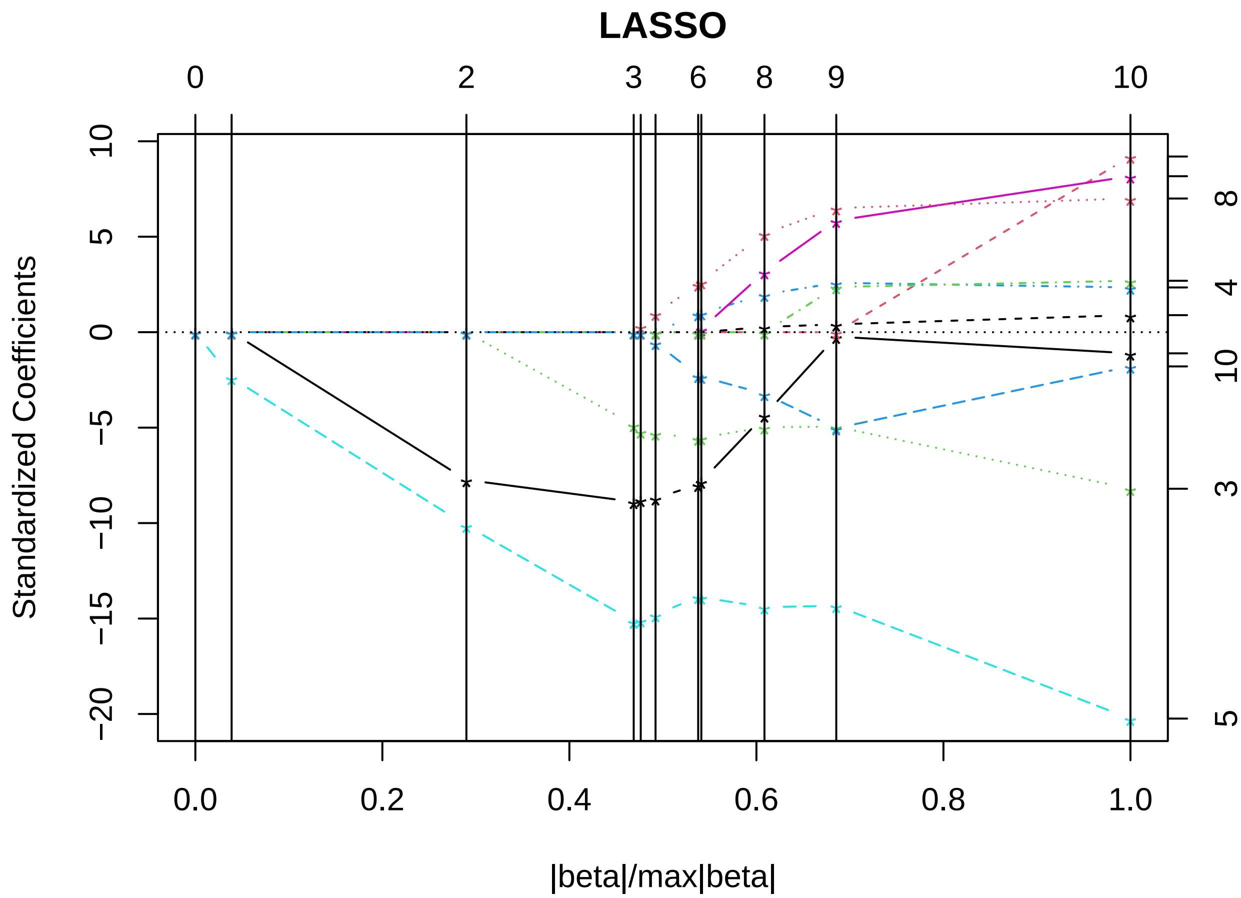


With the exception of drat and am, every variable has very high VIF, indicating that the variance of all the $\hat{\beta}_i$ are high owing to multi-collinearity.

To select the most important variables, we use two approaches: LASSO and Stepwise Forward Selection.

```
##
##
##                              Stepwise Summary
## -------------------------------------------------------------------------
## Step    Variable        AIC        SBC       SBIC        R2       Adj. R2
## -------------------------------------------------------------------------
##  0      Base Model    208.756    211.687    115.061    0.00000    0.00000
##  1      wt            166.029    170.427     74.373    0.75283    0.74459
##  2      cyl           156.010    161.873     66.190    0.83023    0.81852
##  3      hp            155.477    162.805     66.696    0.84315    0.82634
## -------------------------------------------------------------------------
##
## Final Model Output
## ------------------
##
##                          Model Summary
## ---------------------------------------------------------------
## R                       0.918       RMSE                 2.349
## R-Squared               0.843       MSE                  5.519
## Adj. R-Squared          0.826       Coef. Var           12.501
## Pred R-Squared          0.796       AIC                155.477
## MAE                     1.845       SBC                162.805
## ---------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##  AIC: Akaike Information Criteria
##  SBC: Schwarz Bayesian Criteria
##
##                                ANOVA
## --------------------------------------------------------------------
##                Sum of
##               Squares        DF    Mean Square      F          Sig.
## --------------------------------------------------------------------
## Regression    949.427         3        316.476    50.171    0.0000
## Residual      176.621        28          6.308
## Total        1126.047        31
## --------------------------------------------------------------------
##
##                                   Parameter Estimates
## ----------------------------------------------------------------------------------------
##       model      Beta    Std. Error    Std. Beta      t        Sig      lower      upper
## ----------------------------------------------------------------------------------------
## (Intercept)    38.752         1.787                 21.687    0.000    35.092     42.412
##          wt    -3.167         0.741       -0.514    -4.276    0.000    -4.684     -1.650
##         cyl    -0.942         0.551       -0.279    -1.709    0.098    -2.070      0.187
##          hp    -0.018         0.012       -0.205    -1.519    0.140    -0.042      0.006
## ----------------------------------------------------------------------------------------
```

Using the Akaike Information Criterion (AIC) as a metric, we add one explanatory variable at a time until the AIC doesn't decrease, this gives us three variables: wt, cyl, and hp.




```
##
## Call:
## lars(x = cars_matrix, y = cars$mpg)
## R-squared: 0.869
## Sequence of LASSO moves:
##      wt cyl hp am carb drat qsec vs gear disp
## Var   5   1  3  8   10    4    6  7    9    2
## Step  1   2  3  4    5    6    7  8    9   10
```

LASSO suggests the same 3 variables in the same order as using Forward Selection with AIC, wt, cyl, and hp. Looking at the LASSO plot, LASSO determines that wt and cyl are relatively more 'important' than all the other variables, since the regularisation has to be quite low for LASSO to include hp (|beta|/max|beta| > 0.2).

Now we build a new model based on these variable selection techniques:

$$\text{mpg} = \beta_0 + \beta_1 \text{wt} + \beta_2 \text{cyl} + \beta_3 \text{hp} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

```
##
## Call:
## lm(formula = mpg ~ wt + cyl + hp, data = cars)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9290 -1.5598 -0.5311  1.1850  5.8986
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
## wt          -3.16697    0.74058  -4.276 0.000199 ***
## cyl         -0.94162    0.55092  -1.709 0.098480 .
## hp          -0.01804    0.01188  -1.519 0.140015
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.512 on 28 degrees of freedom
## Multiple R-squared:  0.8431, Adjusted R-squared:  0.8263
## F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

The p-value for the F-test is virtually zero, indicating that the explanatory variables do have a linear relationship with **mpg**. $R^2 = 0.8431$ indicates that about 84.31 percent of the variance in **mpg** can be explained by the explanatory variables. However, only wt is significant. Our goal is to succinctly summarise the relationship between **mpg** and other explanatory variables, ideally, we would like all variables to be significant. The obvious variable to remove is hp, since it is the third suggested variable by both LASSO and Forward selection. So we now build the model:

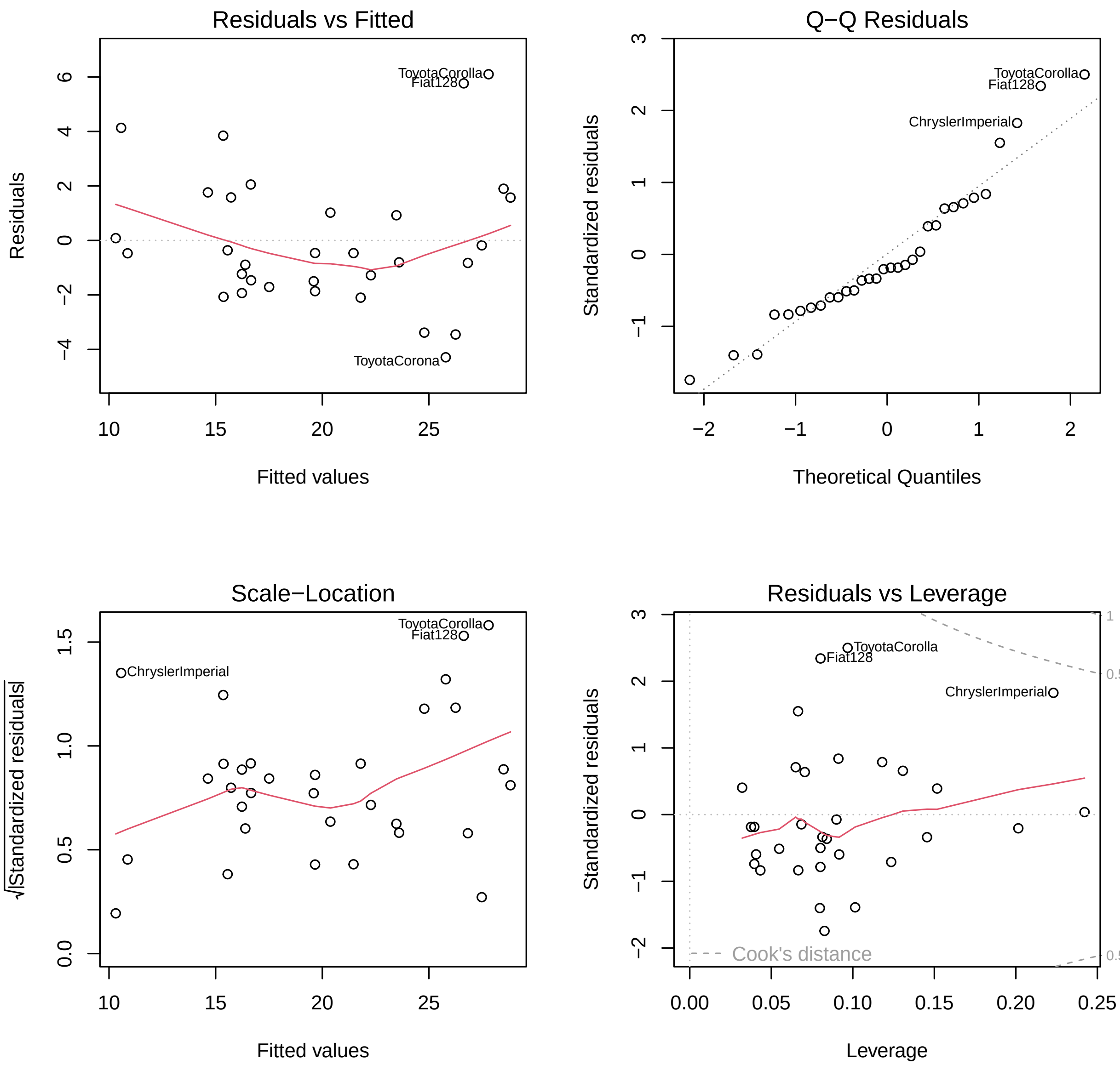
$$\text{mpg} = \beta_0 + \beta_1 \text{wt} + \beta_2 \text{cyl} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

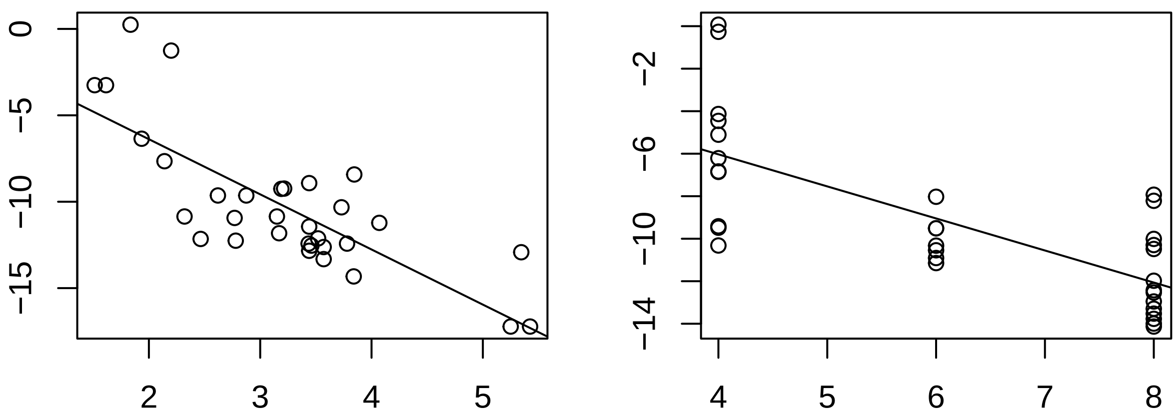
The p-value for the F-test is virtually zero, indicating that the explanatory variables do have a linear relationship with **mpg**. $R^2 = 0.8302$ indicates that about 83.32 percent of the variance in **mpg** can be explained by the explanatory variables. Both wt and cyl are significant explanatory variables. The R^2 value of 0.8302 is not that much lower than the R^2 for the model with ALL the variables ($R^2 = 0.869$), so most of the variance in **mpg** that could be explained by ALL the variables could be explained by just wt and cyl, indicating that these variables are good choices for the explanatory variables.

Now we look at the residuals from our model:

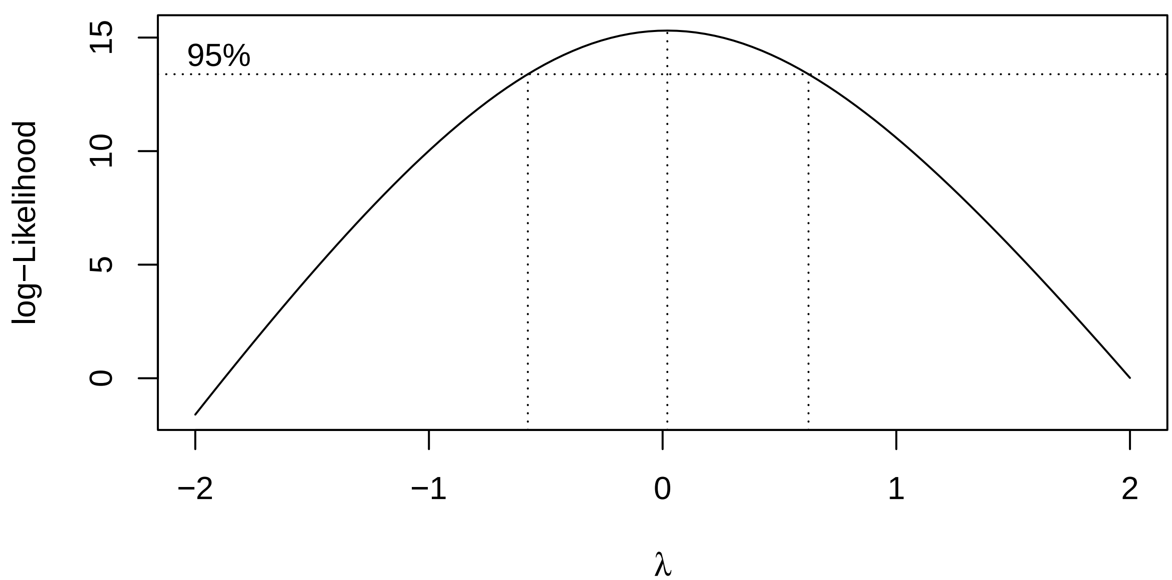


The residuals-fitted values exhibits U-shape curvature, indicating that there is non-linearity, hence. The Normal Q-Q plot shows moderate deviation from the diagonal line, this means the residuals are not normally distributed. We ought to transform either the response or the explanatory variable (or both) to deal with both problems.

These are partial residual plots for wt and cyl:



Neither shows severe curvature, so we do not add higher powers of cyl or wt. Next, we consider using a Box-Cox transformation for the response variable:



This curve shows the log-likelihood of the data as a function of λ, the 95% confidence interval for λ is nearly centred at 0 and 1 is well outside the 95% confidence interval. This clearly indicates that taking logarithm of **mpg** would make the distribution of the residuals more normal.

So we build a new model:

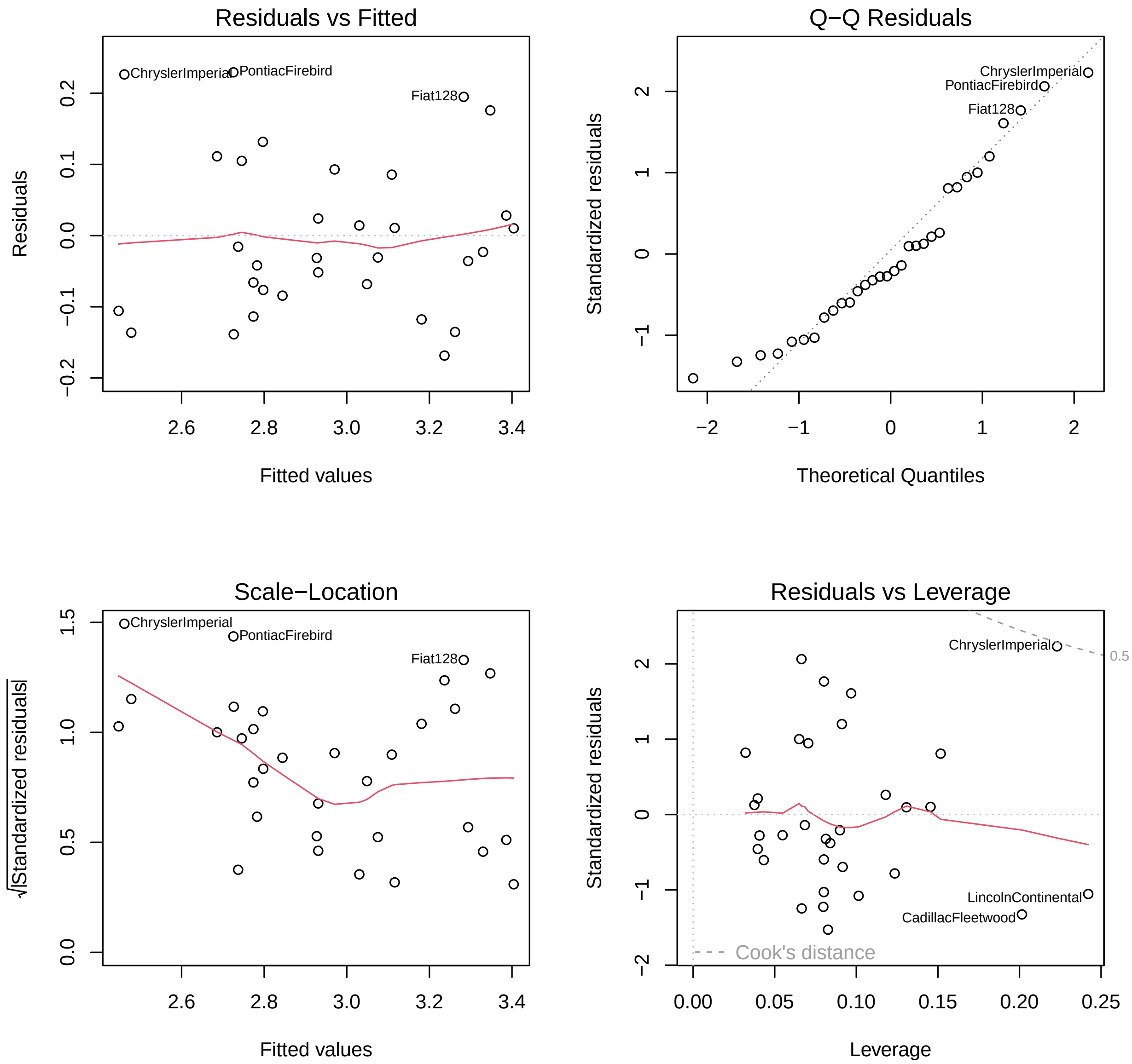
$$\log(\text{mpg}) = \beta_0 + \beta_1 \text{wt} + \beta_2 \text{cyl} + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

```
##
## Call:
## lm(formula = log(mpg) ~ cyl + wt, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16853 -0.07837 -0.02685  0.08745  0.22941
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.93867    0.07688  51.230  < 2e-16 ***
## cyl         -0.06704    0.01859  -3.606  0.00115 **
## wt          -0.17604    0.03393  -5.188  1.5e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1151 on 29 degrees of freedom
## Multiple R-squared:  0.8602, Adjusted R-squared:  0.8506
## F-statistic: 89.24 on 2 and 29 DF,  p-value: 4.061e-13
```

The p-value for the F-test is virtually zero, indicating that the explanatory variables do have a linear relationship with **log(mpg)**. $R^2 = 0.8602$ indicates that about 86.02 percent of the variance in **log(mpg)** can be explained by the explanatory variables. Both wt and cyl are significant explanatory variables.

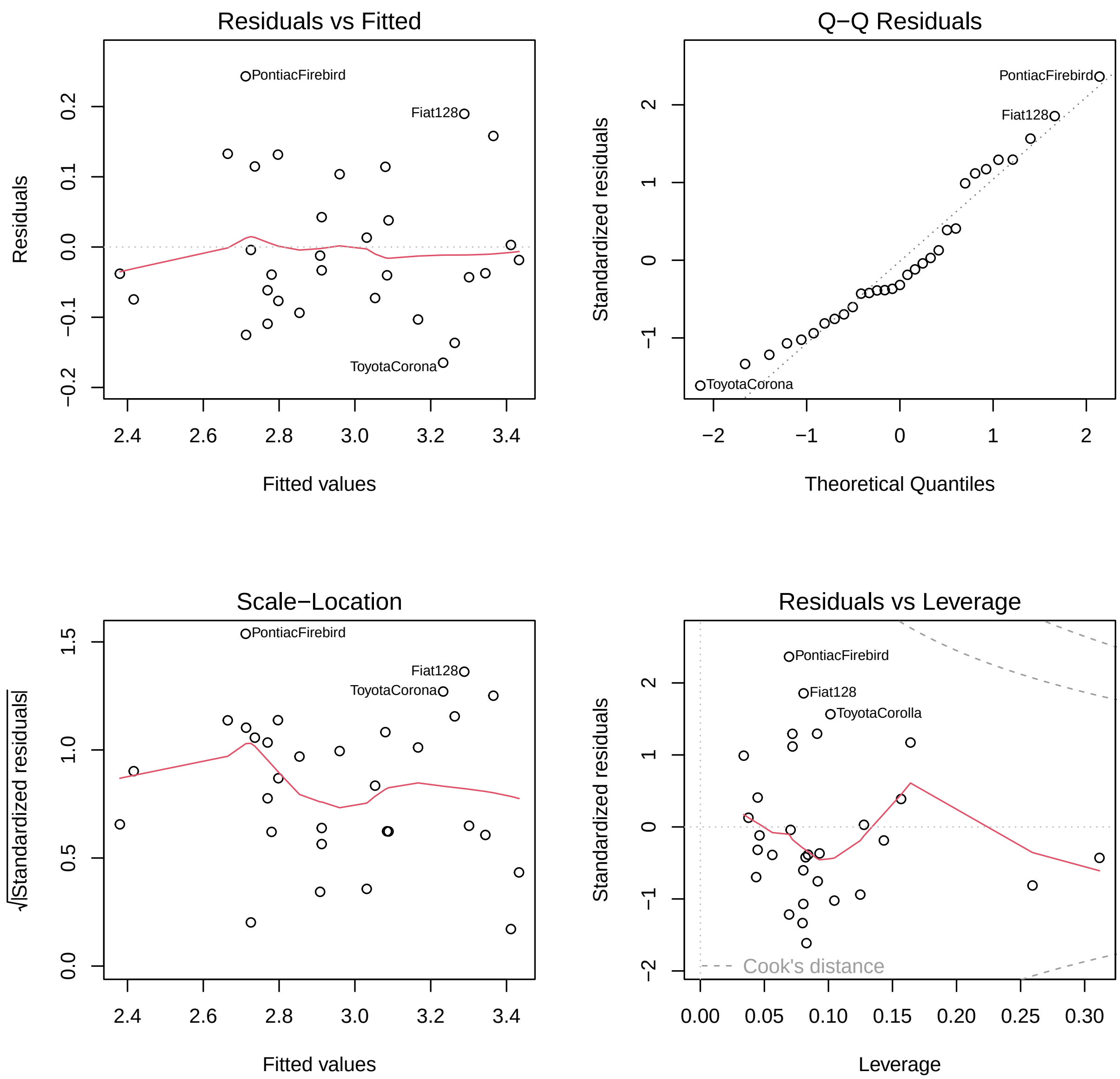
Now we look at the model's residuals:



Looking at the residual-fitted value plot, we can see that there is minimal curvature and there does not seem to be heteroskedasticity. The normal Q-Q however, shows that the residuals are still not perfectly normal. Looking at the residuals-leverage plot, one point we should be concerned about is ‘Chrysler Imperial’, its Cook’s distance is substantially higher than other data points, hence we say that it is an influential observation and remove it from our data set. Now we refit the same model without ‘Chrysler Imperial’.

```
##
## Call:
## lm(formula = log(mpg) ~ wt + cyl, data = cars)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.16476 -0.07363 -0.03320  0.07310  0.24309
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.98193    0.07344  54.222  < 2e-16 ***
## wt          -0.21025    0.03449  -6.097 1.41e-06 ***
## cyl         -0.05771    0.01765  -3.270  0.00285 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1066 on 28 degrees of freedom
## Multiple R-squared:  0.881,  Adjusted R-squared:  0.8725
## F-statistic: 103.6 on 2 and 28 DF,  p-value: 1.145e-13
```



Now the model has an even higher R^2, 0.881. The p-value for the F-test is still virtually zero. Both wt and cyl are significant explanatory variables. The residuals vs fitted plot shows minimal curvature and heteroskedesticity. The Normal Q-Q plot shows that the residuals is approximately normally distributed.

Therefore, our final model is:

$$\log(\text{mpg}) = \beta_0 + \beta_1 \text{wt} + \beta_2 \text{cyl} + \epsilon$$

i.e.:

$$\text{mpg} = \alpha_0 \cdot \alpha_1^{\text{wt}} \cdot \alpha_2^{\text{cyl}} \times e^{\epsilon},$$

where $\epsilon \sim N(0, \sigma^2)$.

Our estimates for the parameters are:

$$\hat{\beta}_0 = 3.98193,\ \hat{\beta}_1 = -0.21025,\ \hat{\beta}_2 = -0.05771,\ \hat{\sigma} = 0.1066$$

or:

$$\hat{\alpha}_0 = 53.62,\ \hat{\alpha}_1 = 0.81,\ \hat{\alpha}_2 = 0.944,\ \hat{\sigma} = 0.1066.$$

Therefore, a unit increase in weight (an increase of 1000 lbs) is associated with $81 - 100 = -19\%$ increase (i.e. 19% decrease) in miles per gallon and a unit increase in cylinder (i.e. one more cylinder) is associated with $0.944 - 100 = -5.6\%$ increase (i.e. 5.6% decrease) in miles per gallon.

$\hat{\sigma}^2 = 0.1066$, $e^{2\times 0.1066} \approx 1.24$, therefore, we can expect about 95% of cars' miles per gallon to be within 24% of the formula $\text{mpg} = \alpha_0 \cdot \alpha_1^{\text{wt}} \cdot \alpha_2^{\text{cyl}}$.

The following table shows some values of **mpg** predicted by our model:

wt	cyl	predictions
1.000	4	34.49544
1.000	6	30.73500
1.000	8	27.38449
2.000	4	27.95452
2.000	6	24.90712
2.000	8	22.19193
3.000	4	22.65387
3.000	6	20.18431
3.000	8	17.98397
4.000	4	18.35832
4.000	6	16.35703
4.000	8	14.57390
5.000	4	14.87727
5.000	6	13.25546
5.000	8	11.81044
5.345	8	10.98410

The last row is the weight and number of cylinder for Chrysler Imperial (which we removed from the model); our model predicts 10.984 miles per gallon for the Chrysler Imperial, which is very much smaller than the real value of 14.7. This perhaps indicates that the Chrysler Imperial was special in some way that made its fuel economy quite high despite its weight and the number of cylinders its engine had.

2.

a.

Job satisfaction is the response variable, workplace environment is the explanatory variable.

b.

Resting heart rate is the response variable, age and exercise frequency are explanatory variables.

c.

BMI is the response variable, age, gender, hours of screen time per day, and physical activity level are explanatory variables.

d.

Annual profit is the response variable, type of business, location, advertising budget and number of employees are explanatory variables.

3.

a.

Generally speaking, cuisine is nominal because there is no sensible way to order them.

b.

Ordinal

c.

Generally speaking, method of payment is nominal because there is no sensible way to order them.

d.

Interval

e.

Ordinal, because the gaps of different SIZEs may not be equal

f.

Generally speaking, preferred vacation type is nominal because there is no sensible way to order them.

g.

Nominal

h.

Generally speaking, weather type is nominal because there is no sensible way to order them; this is because 'weather' encompasses a wide variety of phenomena, including temperature, humidity, wind speed, and sunshine, so it is not easy to order them in a sensible way.

made by 蘇羿豪、馬康麟 助教