Analysis of transmission type influence on MPG

Summary

In this report, `mtcars` data set was analysed in order to explore the relationship between a set of variables and miles per gallon (MPG). In particular, influence of transmission type (automatic vs manual) was considered. The analysis was performed by fitting linear models for MPG. Analysis showed that there is a significant difference in MPG between cars with automatic and manual transmissions. However, the best linear model for MPG as outcome did not include transmission type as regressor. More specifically, having car weight and number of cylinders fixed, transmission type does not influence MPG. [Report was prepared with `knitr`. Source code may be found at `https://github.com/vbalys/coursera-regression-project`].

Data

The `mtcars` data set consists of 32 observations of 11 variables. `mpg` is miles per galon - outcome that we are modeling, and `am` is transmission type (`0` - automatic, `1` - manual) - variable of our interest. The remaining 9 variables are possible regressors for a model explaining `mpg` values. Fig. 1 in the Appendix summarises dependencies between variables.

Boxplots in Fig. 2 suggest that there is a difference in MPG between transmission types. Indeed, t test confirms that difference between means (24.39 vs 17.15) is significant (`p = 0.001374`). However, this does not necessarily mean that difference in MPG is actually related to transmission type. It is possible that we observe a result of confounding factors that are correlated both with MPG and transmission type. To answer the questions of this research, we have to build linear model for MPG. And only then we will find out if there is a way to quantify influence of transmission type.

Analysis

Let us first start with a simplistic model where MPG is explained only by transmission type (`mpg ~ am`). It is immediately obvious that this model is underspecified. Adjusted R-squared (R^2_{adj}) value is only 0.33, while residuals vs. fitted values plot in Fig. 3 does not show the expected normality of residuals. This comes as no surprise as transmission type is a factor taking only two values, therefore model can predict only two values (means of MPG for each transmission type).

Another candidate is a model that uses all variables as predictors (`mpg ~ .`). In this case, we get overspecified model. Even though $R^2_{adj} = 0.8$ and residuals vs. fitted values plot in Fig. 4 is much better, none of the coefficients are significant. Clearly, we have included too much predictors that in various ways correlate with each other and thus cancel each other out.

```
##                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
```

Now we have to find a model in between these two extreme ones that would reasonably explain variation of MPG, have significant coefficients and approximately normal residuals. We are including transmission type `am` as predictor by default, because we are looking for its influence. For the remaining 9 variables there are 512 (2^9) combinations of including/excluding any of them - clearly too much to check all of them.

Let us first start with a "reasonable" model which we will then try to improve. If we think from a mechanistical point of view, we can come up with some clear candidates for important predictors. Weight (`wt`) is obviously a factor to consider when talking about MPG. Acceleration (`qsec`) is itself an outcome of car design choices, and using it to predict MPG would be not

logical. Similarly, horsepower (`hp`) is a result of other factors, so we choose to not include it in the model. Number of cylinders (`cyl`) and engine displacement (`disp`) are highly related (indeed, correlation is 0.9) to each other and probably interchangeable, therefore we choose to include only one of them - `cyl` which is factor variable with three levels. The remaining ones (rear axle ratio `drat`, engine type `vs`, number of forward gears `gear` and number of carburetors `carb`) are rather obscure ones, therefore we leave them out from the model. So, we start with initial model `mpg ~ am + wt + cyl`.

The resulting linear regression model has significant coefficients for `wt` and `cyl`, $R^2_{adj} = 0.81$, and residuals vs fitted values plot in Fig. 4 looks pretty good with no obvious patterns or assymetry, but with a couple of possible outliers. As coefficient for `am` is not significant ($p = 0.89$), we run nested model tests with ANOVA to check which of variables are needed.

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + cyl
## Model 3: mpg ~ wt + cyl + am
##   Res.Df    RSS Df Sum of Sq       F Pr(>F)
## 1     30 278.32
## 2     29 191.17  1    87.150 12.7728 0.0013 **
## 3     28 191.05  1     0.125  0.0183 0.8933
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We get that we improve model by adding `cyl` variable, but adding `am` does not yield significant improvement. Therefore, we get the final model `mpg ~ wt + cyl`.

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

Finally, we run some diagnostics for the final model (see the last figure in the Appendix). Both from diagnostic plots and calculated `PRESS = resid / (1 - hatvalues)` values we see that `Toyota Corolla`, `Fiat 128` and `Chrysler Imperial` are outliers. This does not mean that there is something wrong with the data, simply MPG for these cars are not correctly explained by our model.

Conclusions

1. There is a significant difference in MPG depending on transmission type. Cars with automatic tranmission have lower MPG than cars with manual transmission.
2. The best linear model for MPG that we managed to come up did not include transmission type as a predictor. This model did include car weight and number of cylinders as regressors. It is highly possible that transmission exerts its influence via car weight.
3. Without linear model with transmission type as a regressor there is no way to quantify its influence on MPG.

Appendix

Fig. 1. Scatterplots for all variable pairs

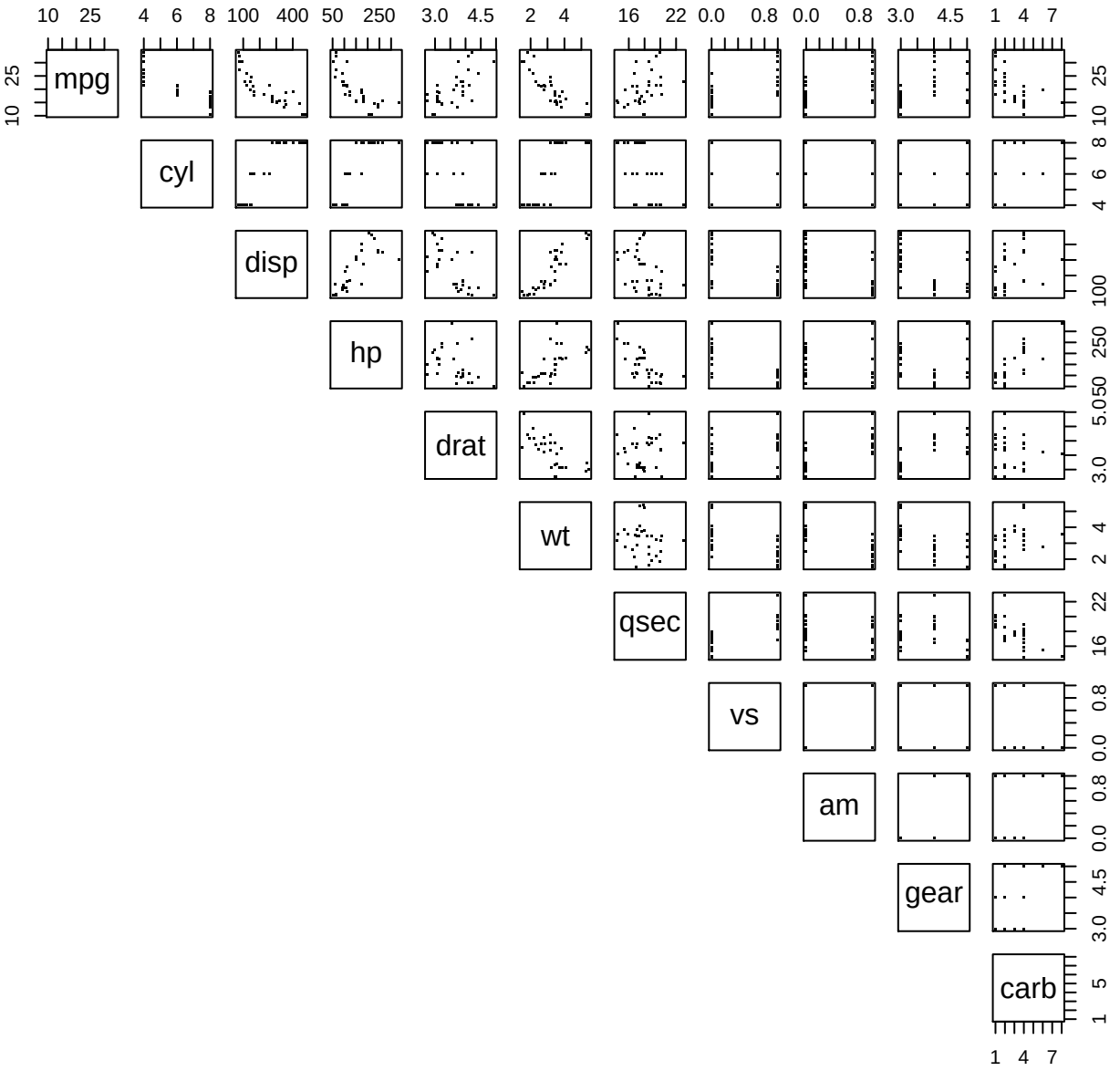


Fig. 2. MPG for transmission types

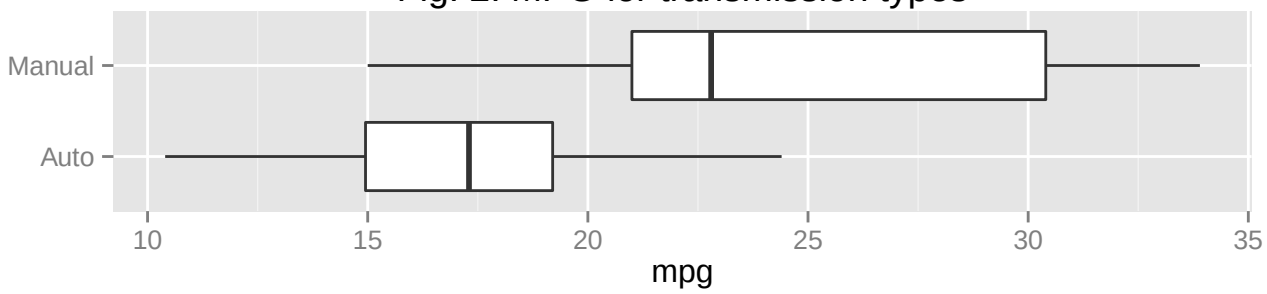


Fig. 3. Res. vs fitted (mpg ~ am)

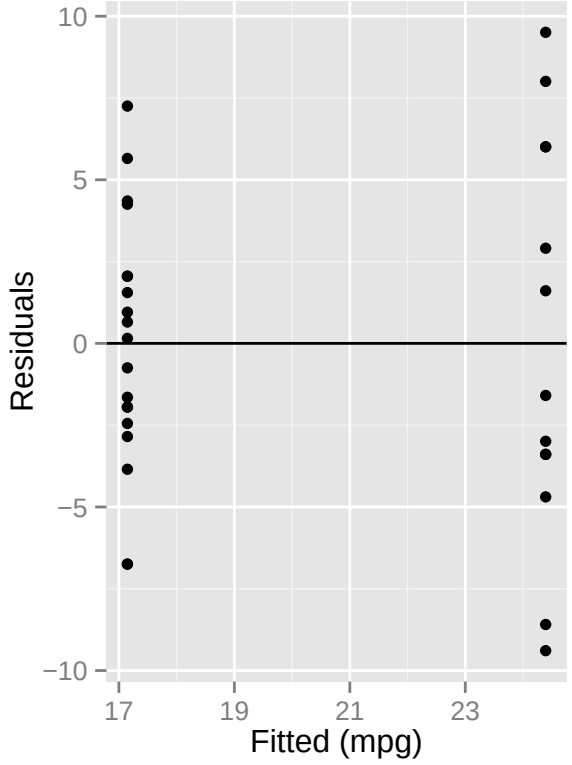


Fig. 4. Res. vs fitted (mpg ~ .)

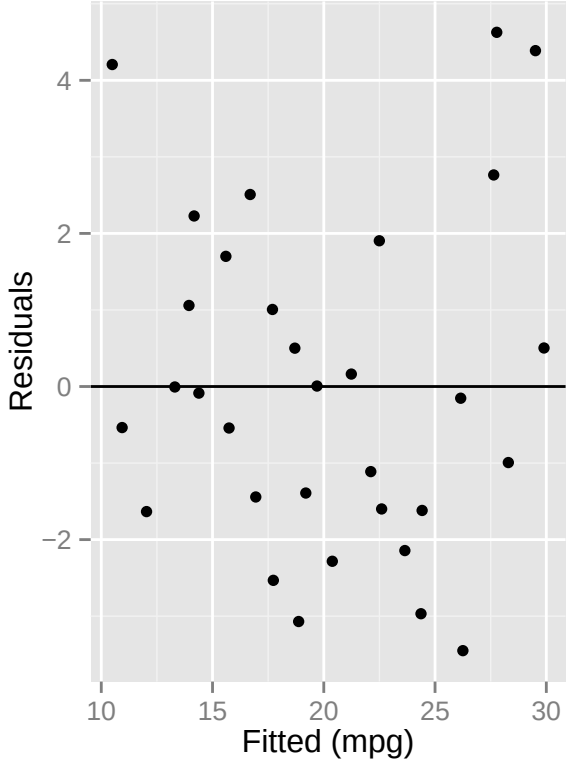
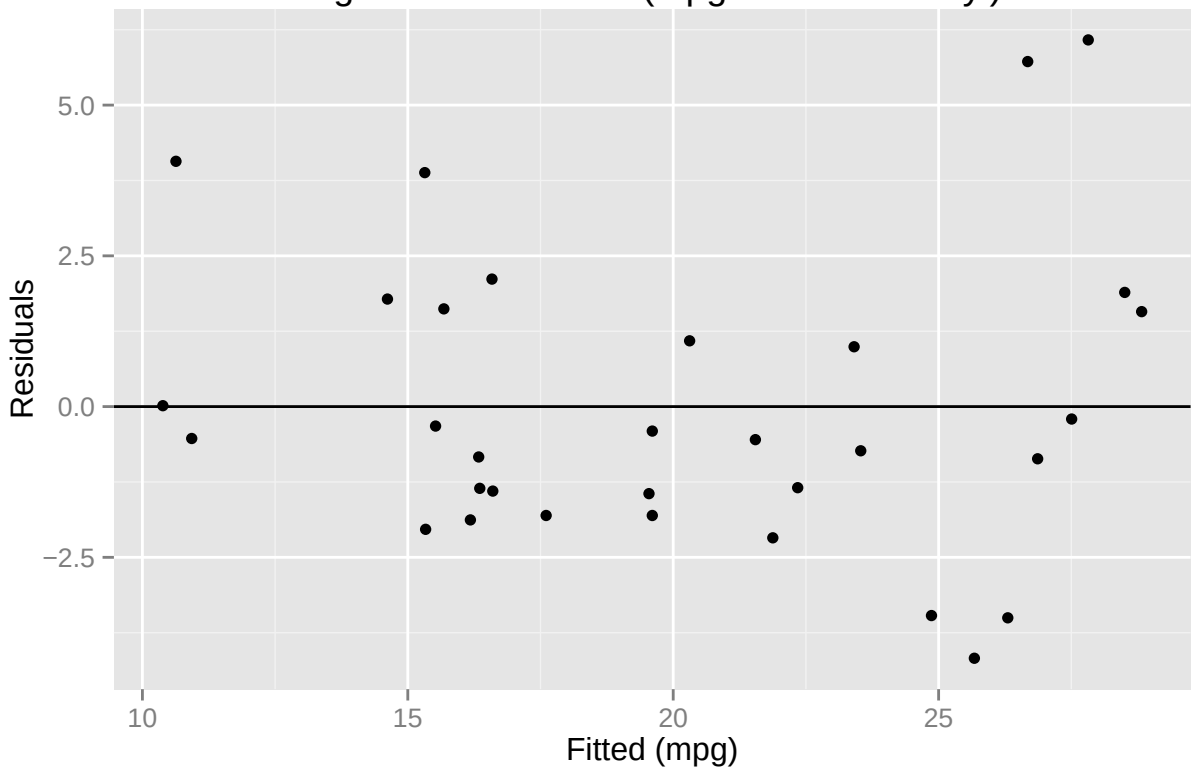
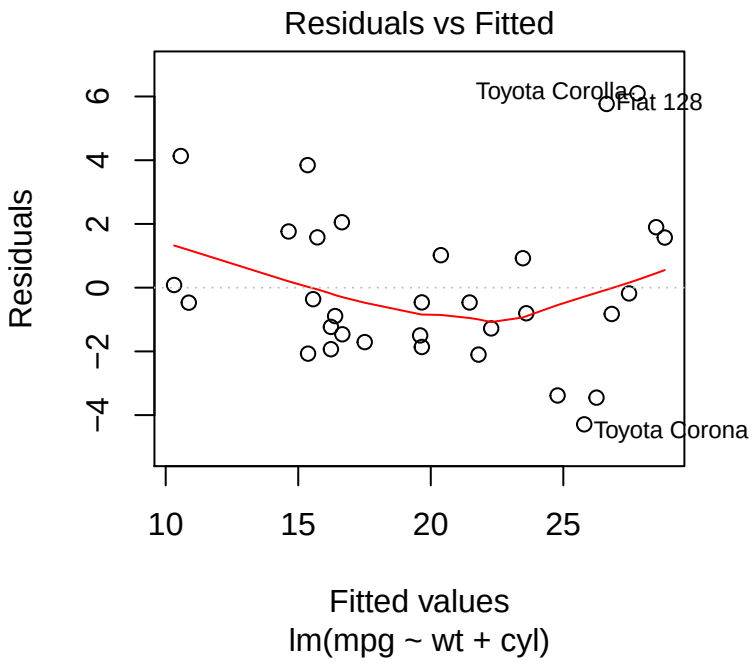
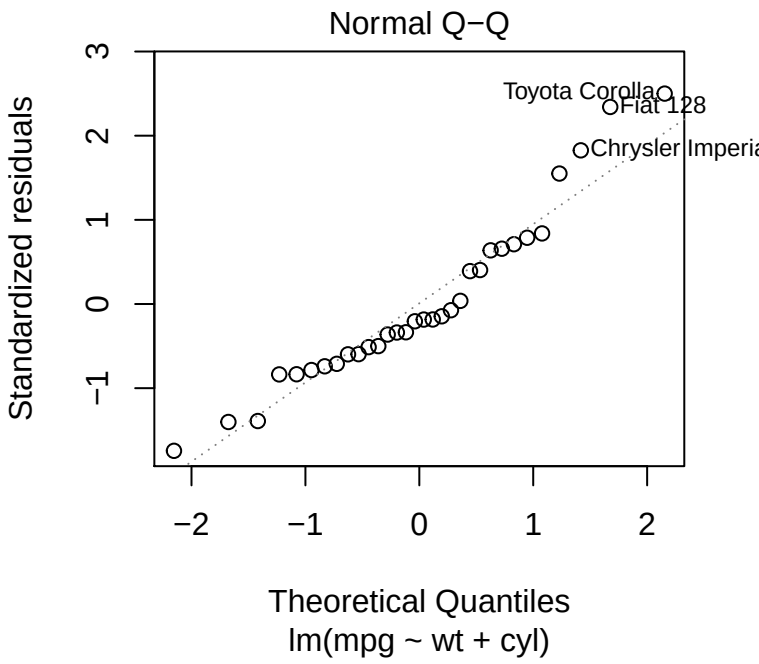
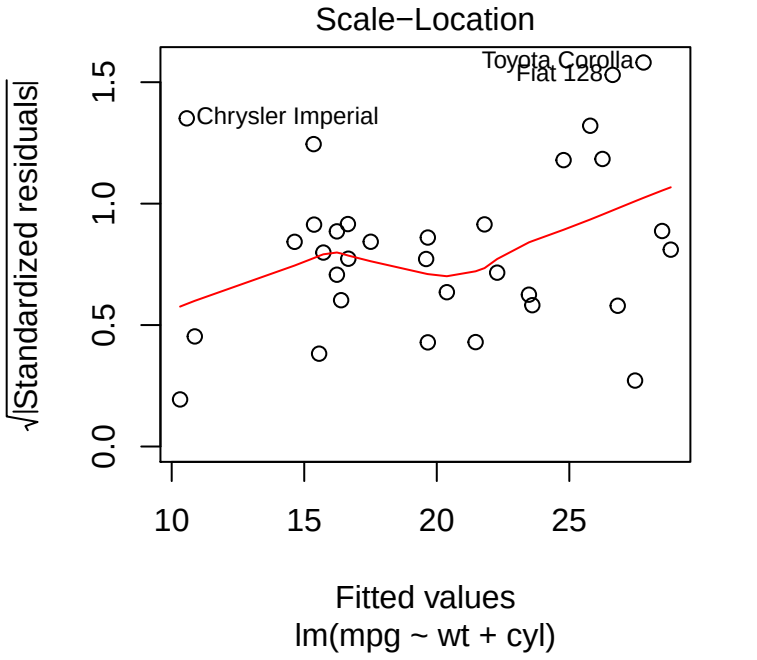


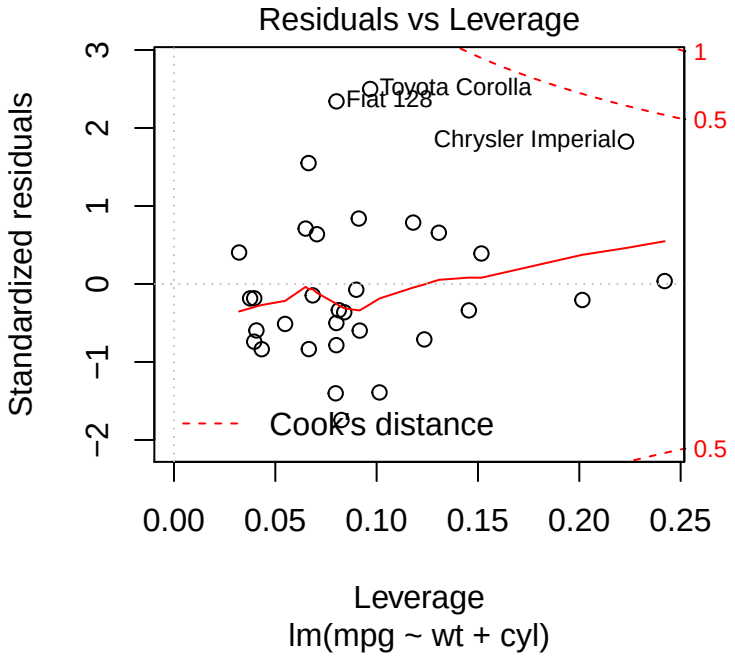
Fig. 5. Res. vs fitted (mpg ~ am + wt + cyl)




Residuals vs Fitted
Residuals
Fitted values
lm(mpg ~ wt + cyl)
Toyota Corolla
Fiat 128
Toyota Corona


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(mpg ~ wt + cyl)
Toyota Corolla
Fiat 128
Chrysler Imperia


Scale−Location
√|Standardized residuals|
Fitted values
lm(mpg ~ wt + cyl)
Toyota Corolla
Fiat 128
Chrysler Imperial


Residuals vs Leverage
Standardized residuals
Leverage
lm(mpg ~ wt + cyl)
Toyota Corolla
Fiat 128
Chrysler Imperial
Cook's distance
1
0.5
0.5