Course Project - Regression Models

Tracy Wilson

Sunday, May 17, 2015

Executive Summary

Our assignment for Motor Trend is to look at the effect of automatic transmissions on fuel efficiency. To do this we will use the mtcars data set that examines the fuel efficency and 10 aspects of automobile design and performance for 32 automobiles (1973 - 1974 models). There are 32 cars in the data set of which 13 have manual transmissions and 19 have automatic transmissions.

."Is an automatic or manual transmission better for MPG" ."Quantify the MPG difference between automatic and manual transmissions"

In this data set on average there is a difference in fuel efficency depending on transmission type such that on average manual vehicles achieve a fuel effiency of 7.2 miles per gallon more than automatic vehicles.

We have found, through this analysis, that transmission type is not a very good predictor of fuel efficiency. By applying analysis of variance (ANOVA) to the dataset, calculating the correlations between the variables, and building a number of models, we were able to identify that the number of cylinders and the weight of the automobile are good predictors of fuel efficiency, achieving an adjusted R squared of 0.82. If we add transmission type to this model, then the difference in fuel effiency for a manual transmission is much smaller, just 0.18 miles per gallon for a vehicle with the same weight and number of cylinders.

Therefore we conclude that number of cylinders and weight are good predictors of fuel efficiency, but transmission type is not.

```r
require(car);
```

```
## Loading required package: car
```

```r
data(mtcars);


#mtcars$cyl  <- factor(mtcars$cyl)
#mtcars$vs   <- factor(mtcars$vs)
#mtcars$gear <- factor(mtcars$gear)
#mtcars$carb <- factor(mtcars$carb)
#mtcars$am   <- factor(mtcars$am,labels=c("Automatic","Manual"))

#help(mtcars)  #opens another web page with help information regarding mtcars data set
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 ...
```

Regression Models and Exploratory Data Analyses

Linear Regression

```r
#Linear Regression
fit <- lm(mpg ~ am + cyl + wt + hp, data=mtcars)

summary(fit)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## wt          -2.60648    0.91984  -2.834   0.0086 **
## hp          -0.02495    0.01365  -1.828   0.0786 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
```

```r
summary(fit)$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 36.14653575 3.10478079 11.642218 4.944804e-12
## am           1.47804771 1.44114927  1.025603 3.141799e-01
## cyl         -0.74515702 0.58278741 -1.278609 2.119166e-01
## wt          -2.60648071 0.91983749 -2.833632 8.603218e-03
## hp          -0.02495106 0.01364614 -1.828433 7.855337e-02
```

```r
data(mtcars)
n <- length(mtcars$mpg)
alpha <- 0.05
fit_limited <- lm(mpg ~ am, data = mtcars)
coef(summary(fit_limited))
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

```
summary(fit_limited)
```

```
##
## 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
```

```
summary(fit_limited)$coefficients
```

```
##             Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am           7.244939   1.764422  4.106127 2.850207e-04
```

Linear regression (heteroskedasticity-robust standard errors)

```
library(lmtest)
```

```
## Loading required package: zoo
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(sandwich)
fit$robse <- vcovHC(fit, type="HC1")
coeftest(fit,fit$robse)
```

```
##
```

```
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 36.146536   2.841426 12.7213 6.452e-13 ***
## am           1.478048   1.393427  1.0607  0.298210
## cyl         -0.745157   0.528924 -1.4088  0.170302
## wt          -2.606481   0.914436 -2.8504  0.008264 **
## hp          -0.024951   0.011004 -2.2675  0.031576 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Predicted values/REsiduals

```
mpg_hat <- fitted(fit)
as.data.frame(mpg_hat)
```

```
##                      mpg_hat
## Mazda RX4           23.58005
## Mazda RX4 Wag       22.91539
## Datsun 710          26.27647
## Hornet 4 Drive      20.55114
## Hornet Sportabout   16.85255
## Valiant             20.03731
## Duster 360          14.76713
## Merc 240D           23.30427
## Merc 230            22.58514
## Merc 280            19.64032
## Merc 280C           19.64032
## Merc 450SE          15.08571
## Merc 450SL          15.97192
## Merc 450SLC         15.84159
## Cadillac Fleetwood  11.38629
## Lincoln Continental 10.68325
## Chrysler Imperial   10.51490
## Fiat 128            27.26293
## Honda Civic         29.13703
## Toyota Corolla      28.23924
## Toyota Corona       24.32068
## Dodge Challenger    17.26781
## AMC Javelin         17.48936
## Camaro Z28          14.06338
## Pontiac Firebird    15.79693
## Fiat X1-9           27.95365
## Porsche 914-2       26.79554
## Lotus Europa        27.88088
## Ford Pantera L      16.81370
## Ferrari Dino        21.56725
## Maserati Bora       13.99959
## Volvo 142E          24.67827
```

```r
mpg_residuals <- residuals(fit)
as.data.frame(mpg_residuals)
```

```
##                     mpg_residuals
## Mazda RX4              -2.5800454
## Mazda RX4 Wag          -1.9153928
## Datsun 710             -3.4764716
## Hornet 4 Drive          0.8488584
## Hornet Sportabout       1.8474494
## Valiant                -1.9373091
## Duster 360             -0.4671339
## Merc 240D               1.0957315
## Merc 230                0.2148572
## Merc 280               -0.4403197
## Merc 280C              -1.8403197
## Merc 450SE              1.3142876
## Merc 450SL              1.3280841
## Merc 450SLC            -0.6415918
## Cadillac Fleetwood     -0.9862887
## Lincoln Continental    -0.2832505
## Chrysler Imperial       4.1851034
## Fiat 128                5.1370721
## Honda Civic             1.2629661
## Toyota Corolla          5.6607556
## Toyota Corona          -2.8206800
## Dodge Challenger       -1.7678086
## AMC Javelin            -2.2893595
## Camaro Z28             -0.7633842
## Pontiac Firebird        3.4030741
## Fiat X1-9              -0.6536453
## Porsche 914-2          -0.7955403
## Lotus Europa            2.5191196
## Ford Pantera L         -1.0137038
## Ferrari Dino           -1.8672544
## Maserati Bora           1.0004137
## Volvo 142E             -3.2782735
```

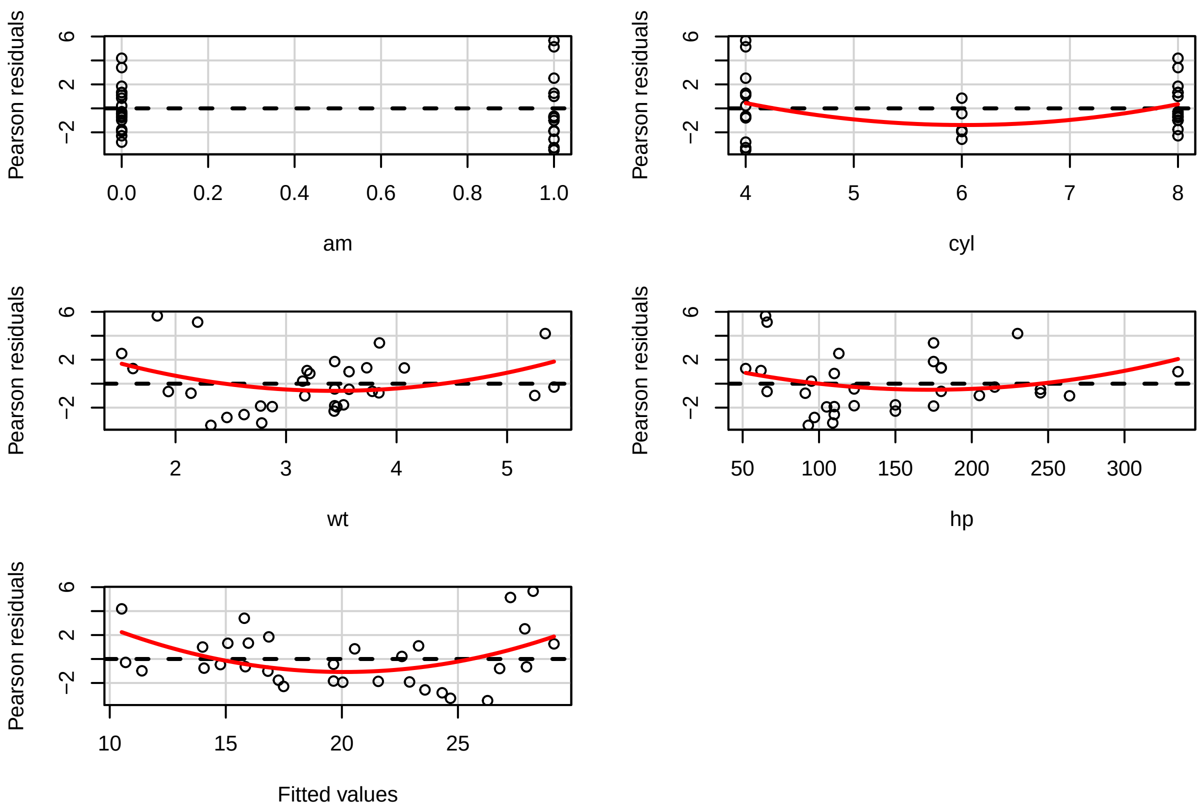
```r
fit2 <-lm(mpg ~ am*(hp + wt), data=mtcars)
summary(fit2)
```

```
##
## Call:
## lm(formula = mpg ~ am * (hp + wt), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9873 -1.4467 -0.5355  1.2614  5.5987
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.70393    2.67515  11.477 1.12e-11 ***
## am          13.74000    4.22337   3.253  0.00316 **
```

```
## hp          -0.04094    0.01363  -3.004  0.00583 **
## wt          -1.85591    0.94511  -1.964  0.06034 .
## am:hp        0.02779    0.01921   1.447  0.15983
## am:wt       -5.76895    2.07201  -2.784  0.00987 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.286 on 26 degrees of freedom
## Multiple R-squared:  0.8793, Adjusted R-squared:  0.8561
## F-statistic: 37.89 on 5 and 26 DF,  p-value: 3.901e-11
```

Diagnostics for linear regression

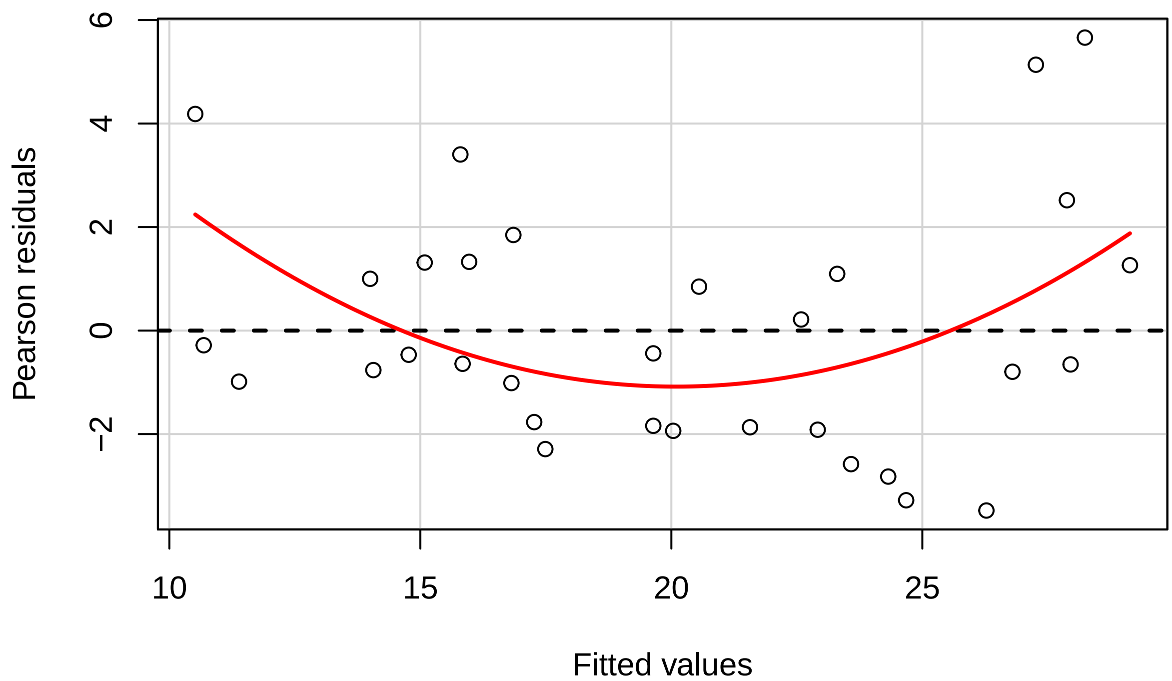
```r
residualPlots(fit)
```




```
##            Test stat Pr(>|t|)
## am             0.627    0.536
## cyl            1.807    0.082
## wt             2.329    0.028
## hp             1.695    0.102
## Tukey test     3.034    0.002
```

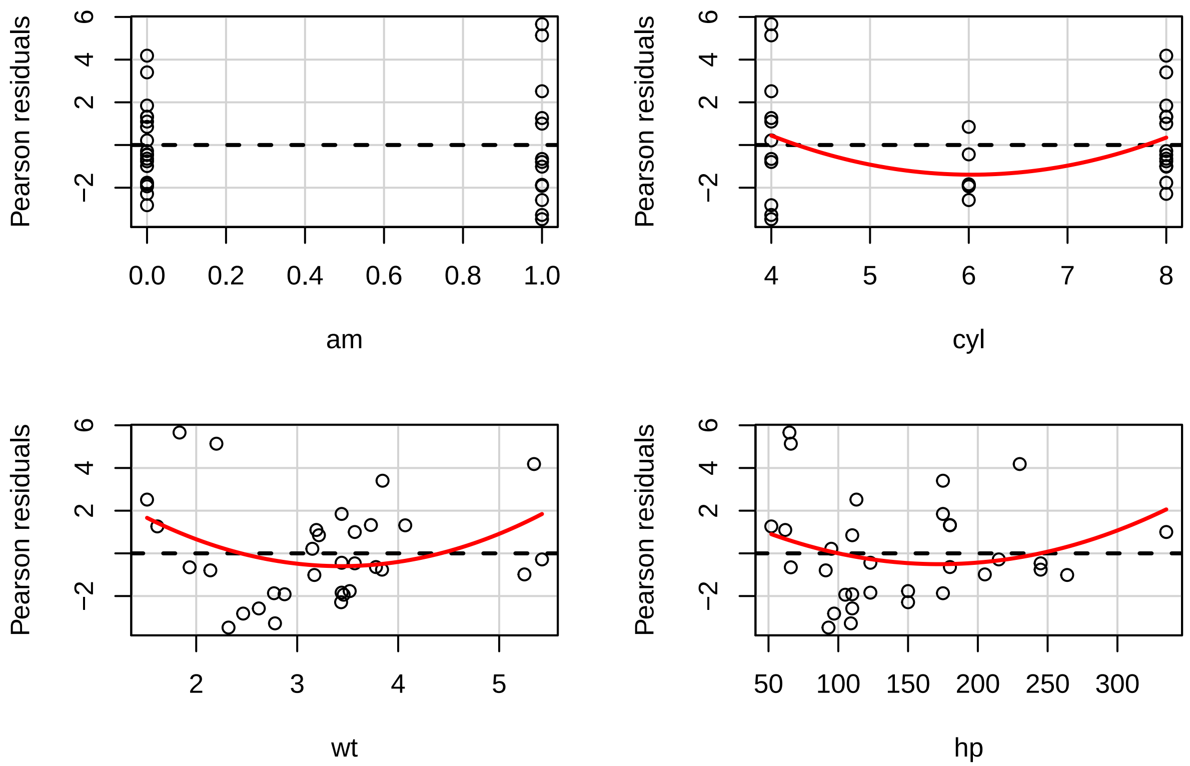
Using'Transmission Type' as is. Variable transmission, cylinder, displacement, and horse power shows some patterns. Other options:

```r
residualPlots(fit, ~ 1, fitted=TRUE) #Residuals vsfitted only
```




```
##            Test stat Pr(>|t|)
## Tukey test     3.034    0.002
```

```r
residualPlots(fit, ~ am + cyl + wt + hp, fitted=FALSE) # Residuals vsam only
```

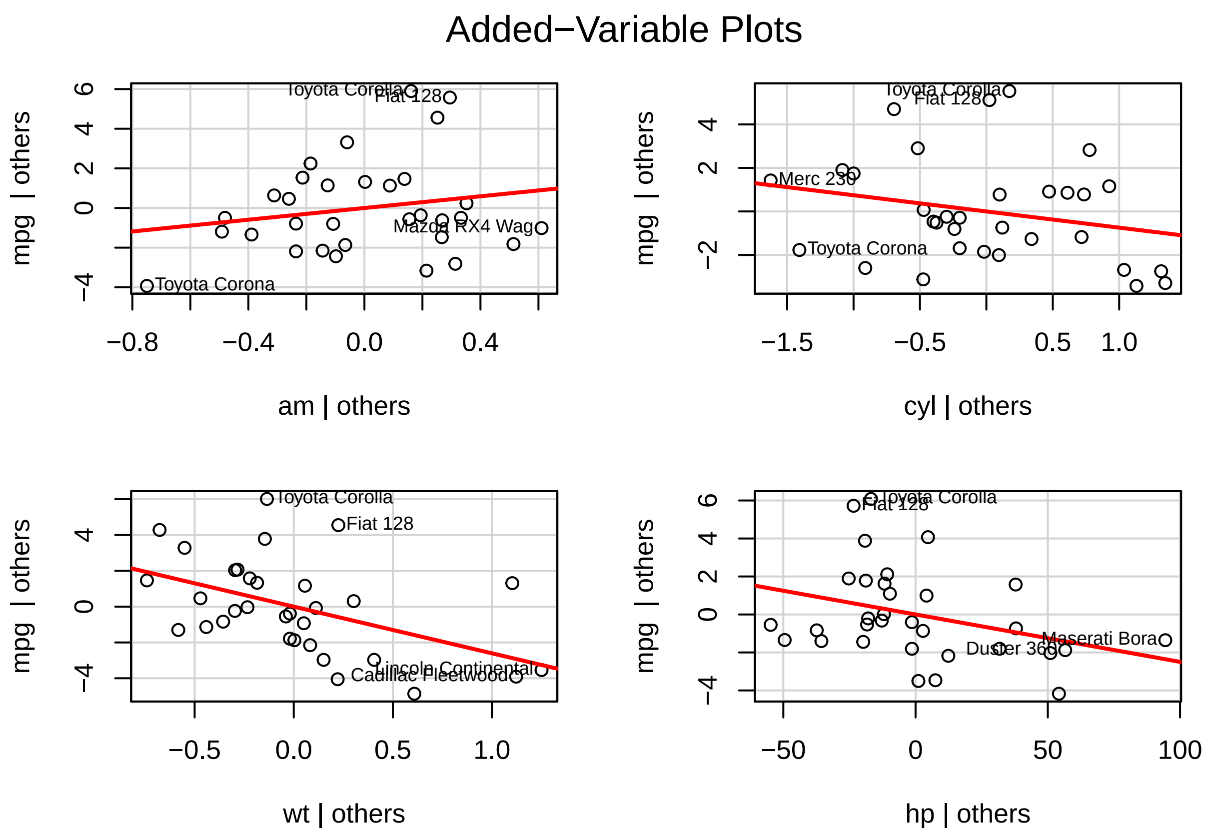



```
##      Test stat Pr(>|t|)
## am       0.627    0.536
## cyl      1.807    0.082
## wt       2.329    0.028
## hp       1.695    0.102
```

What to look for: No patterns,no problems. All p'sshould be non-significant. Modelok if residuals have mean=0 and variance=1 (Fox,316) Tukeytest nullhypothesis: model is additive.

Influential variables-Added-variableplots

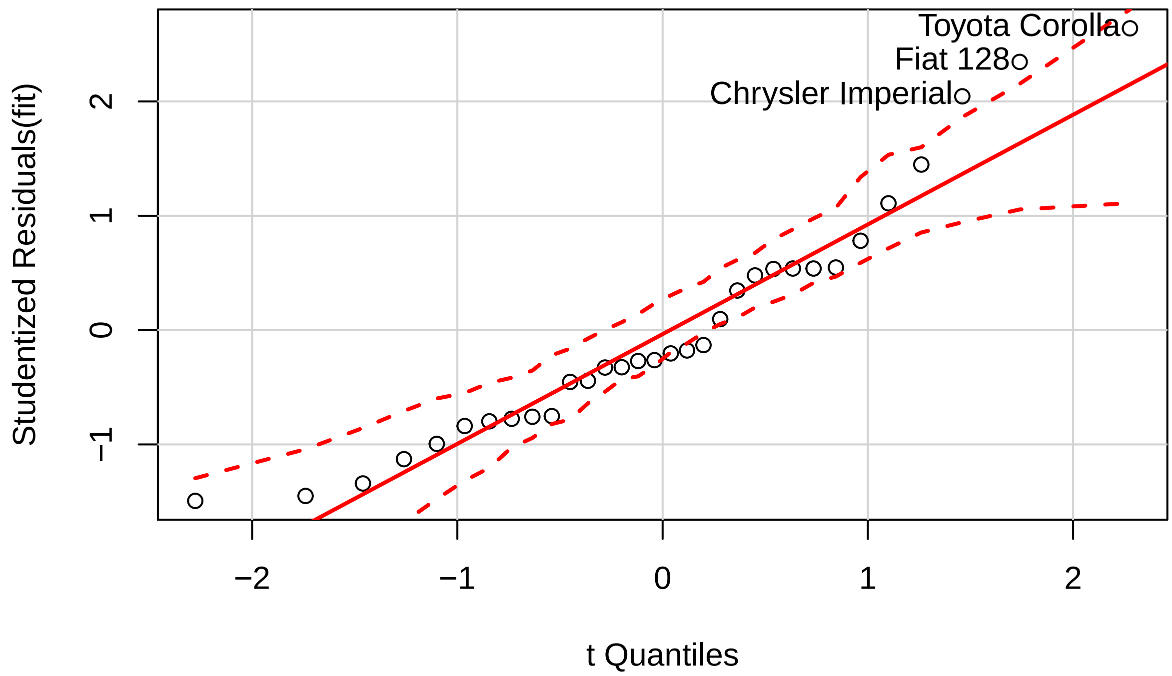
```
avPlots(fit, id.n=2, id.cex=0.7)
```



id.n-id most influential observation id.cex -font size for id. Graphs outcomevspredictor variables holding the rest constant (also called partial-regression plots) Help identify the

Outliers -QQ-Plots

```
qqPlot(fit, id.n=3)
```




```
## Chrysler Imperial          Fiat 128    Toyota Corolla
##                30                31                32
```

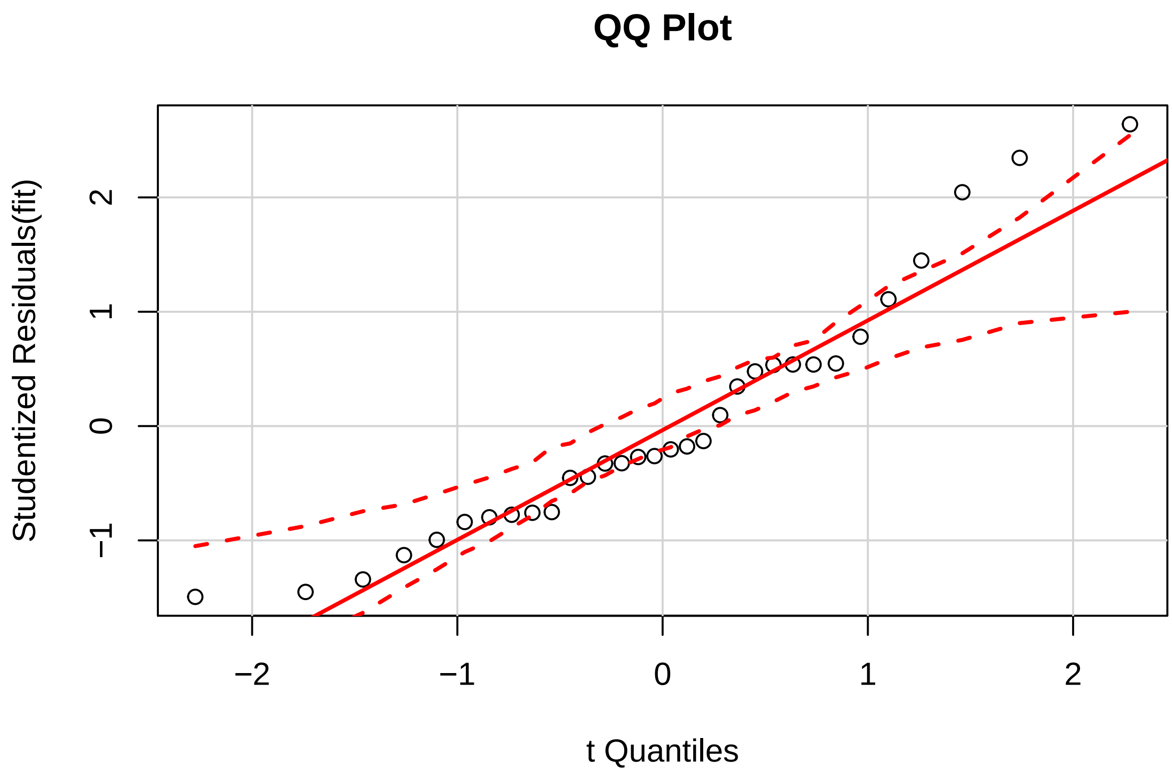
id.n-id observations with high residuals

Outliers -Bonferonni Test

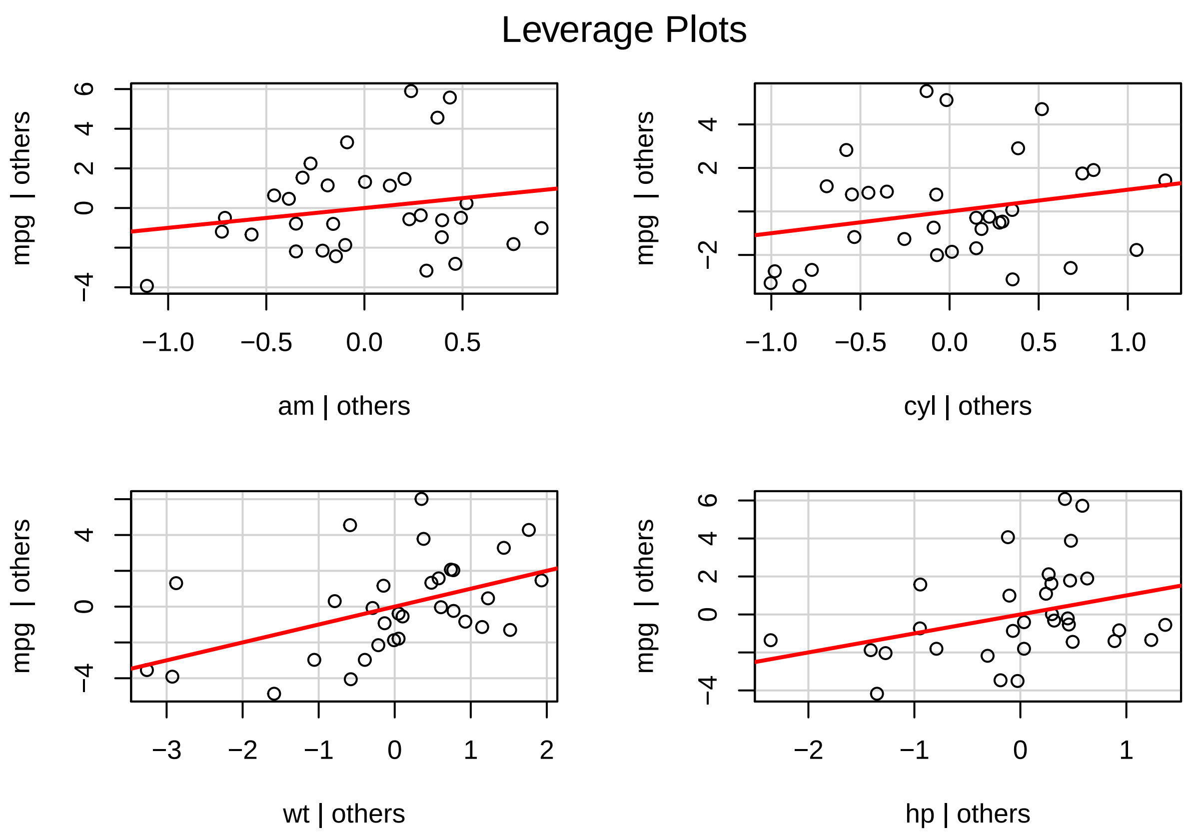
```r
outlierTest(fit)  # Bonferonni p-value for most extreme obs
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##                rstudent unadjusted p-value Bonferonni p
## Toyota Corolla 2.639691           0.013842      0.44293
```

```r
qqPlot(fit, main="QQ Plot") #qq plot for studentized resid
```



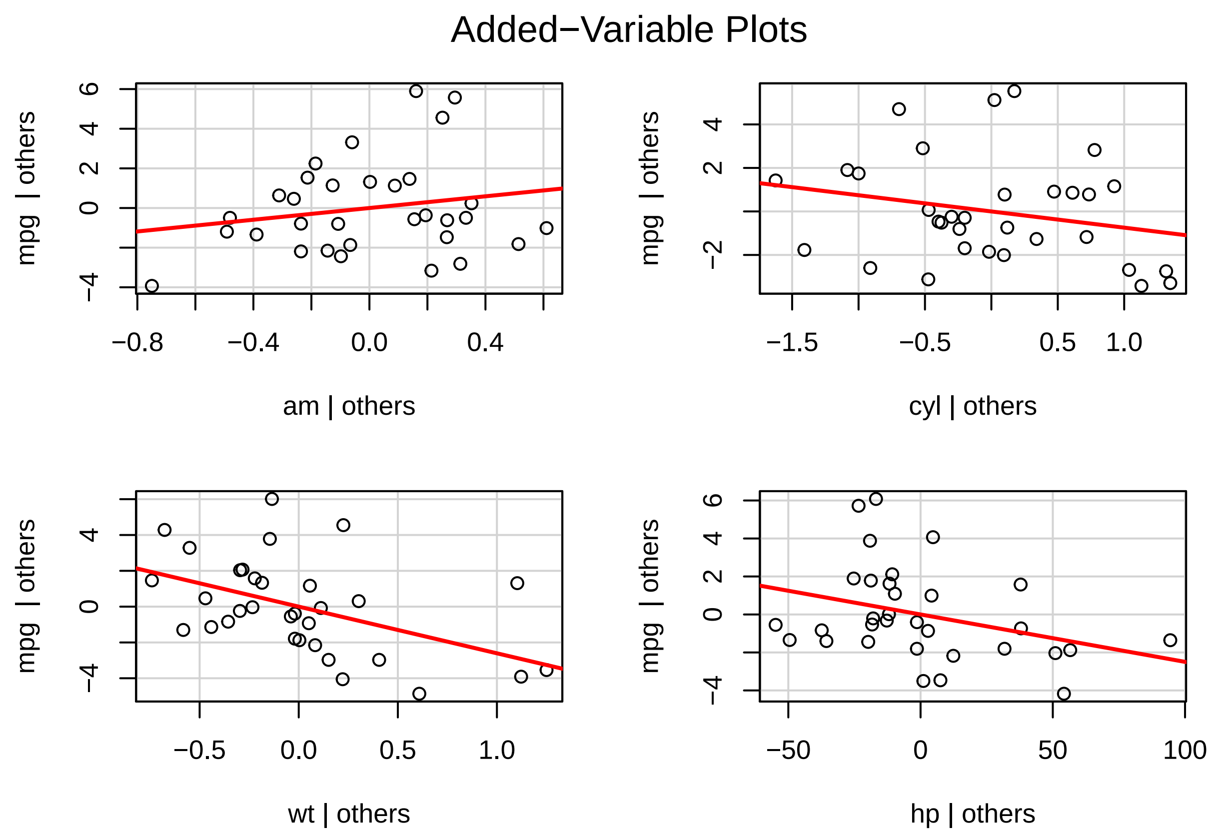

```r
leveragePlots(fit) # leverage plots
```



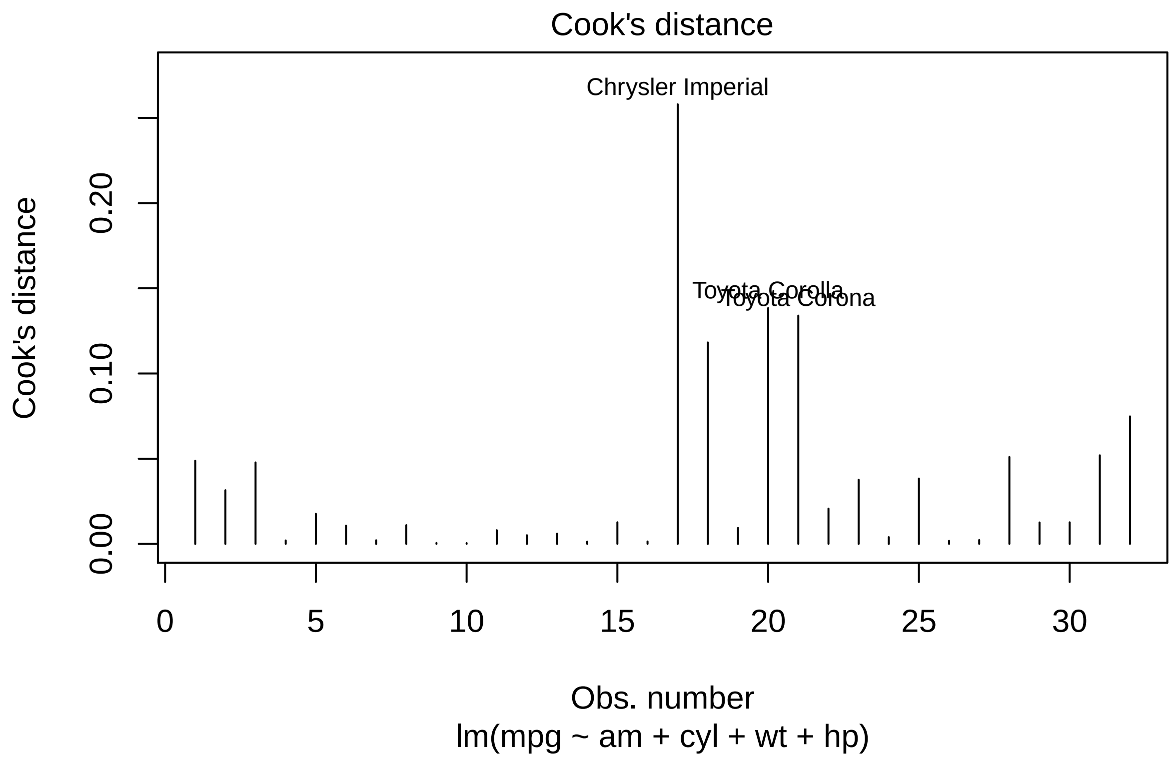
Null for the Bonferonni adjusted outlier test is the observation is an outlier. Here observation related to 'Toyoto Corolla' is an outlier.

Influential Observations

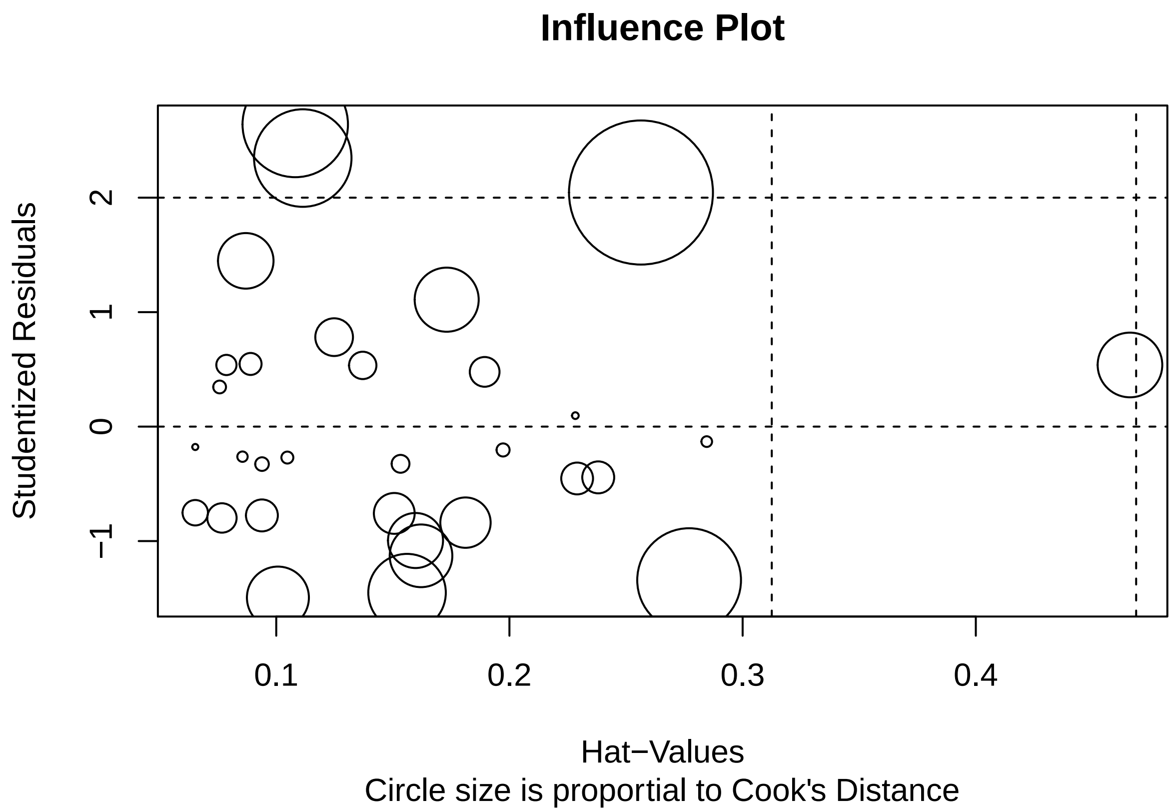
```r
avPlots(fit)
```




```
# Cook's D plot
# identify D values > 4/(n-k-1)
cutoff <- 4/((nrow(mtcars)-length(fit$coefficients)-2))
plot(fit, which=4, cook.levels=cutoff)
```



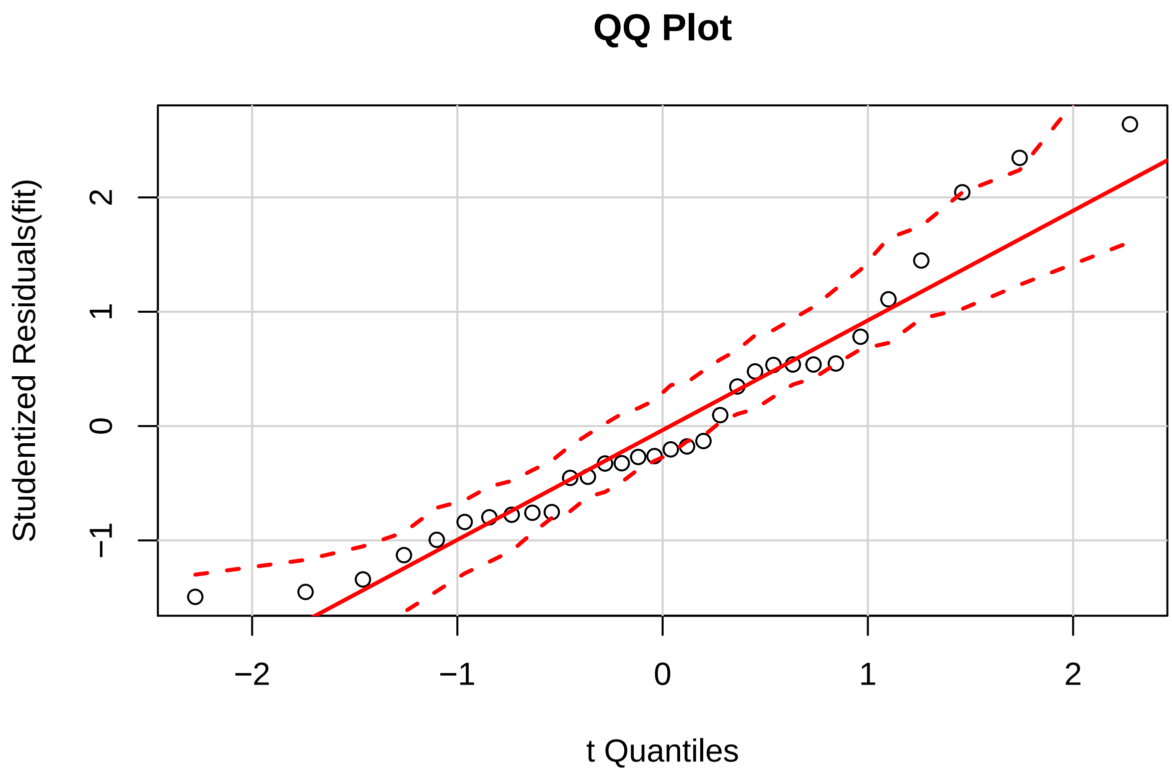

```
# Influence Plot
influencePlot(fit, id.method="identify", main="Influence Plot", sub="Circle size is proportial to Cook
```



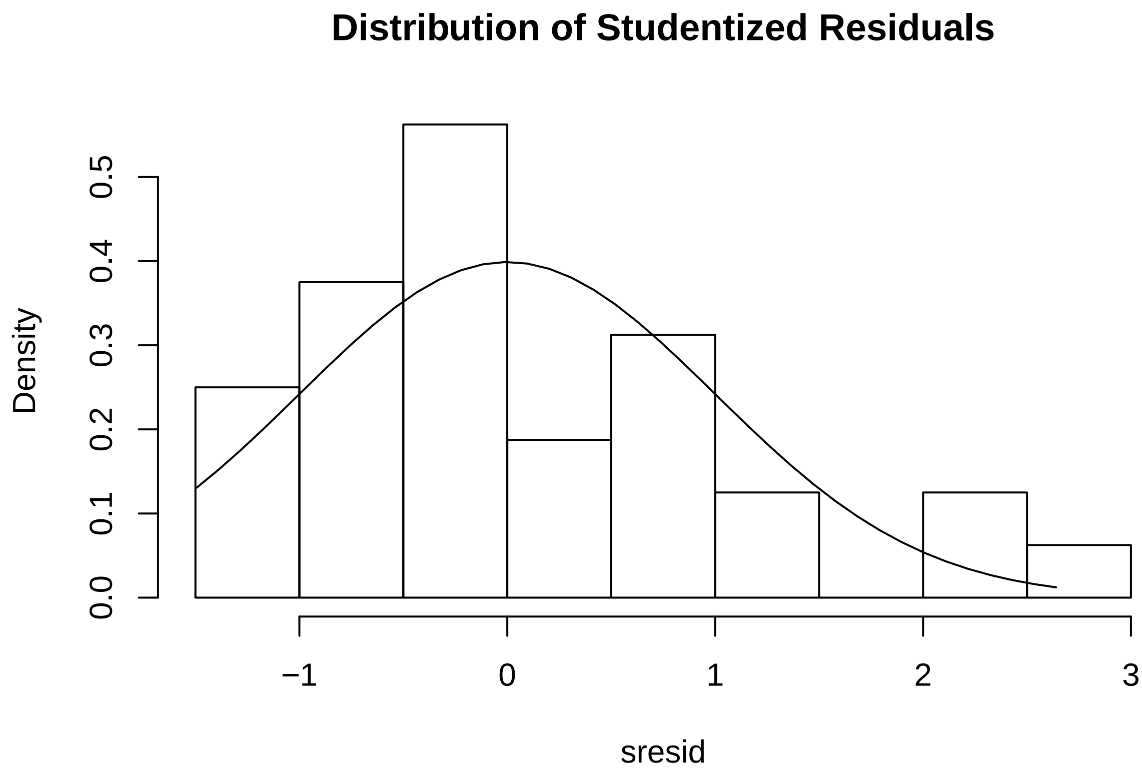
Non-normality

Normality of Residuals

```
# qq plot for studentized resid
qqPlot(fit, main="QQ Plot")
```

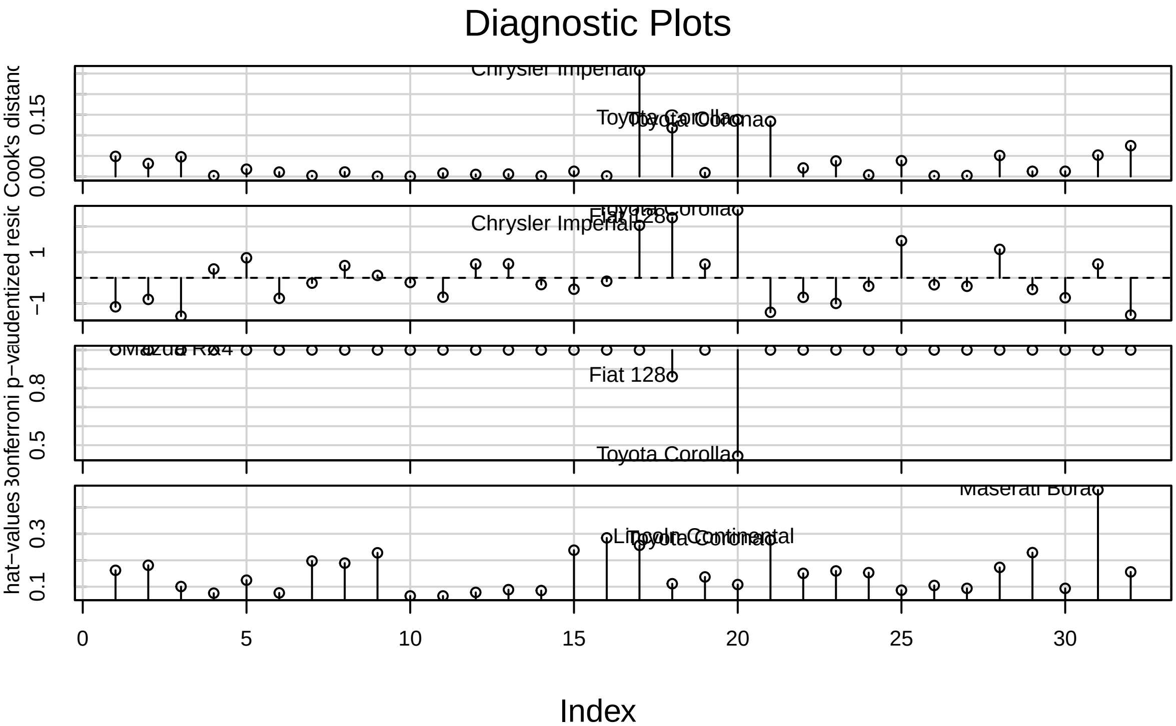



```
# distribution of studentized residuals
library(MASS)
sresid <- studres(fit)
hist(sresid, freq=FALSE,
   main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



High leverage (hat) points

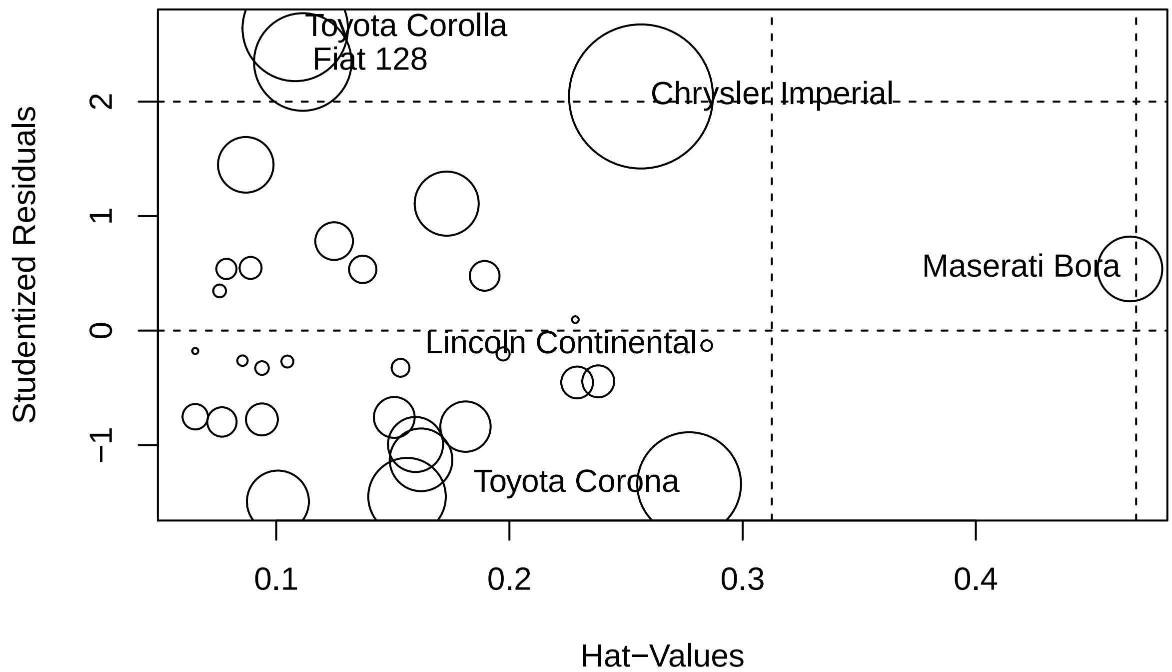
```
influenceIndexPlot(fit, id.n=3)
```



Cook's distance measures how much an observation influences the overall model or predicted values Studentizidedresiduals are the residuals divided by their estimated standard deviation as a way to standardized Bonferronitest to identify outliers Hat-points identify influential observations (have a high impact on the predictor variables)

Influence Plots

```
influencePlot(fit, id.n=3)
```

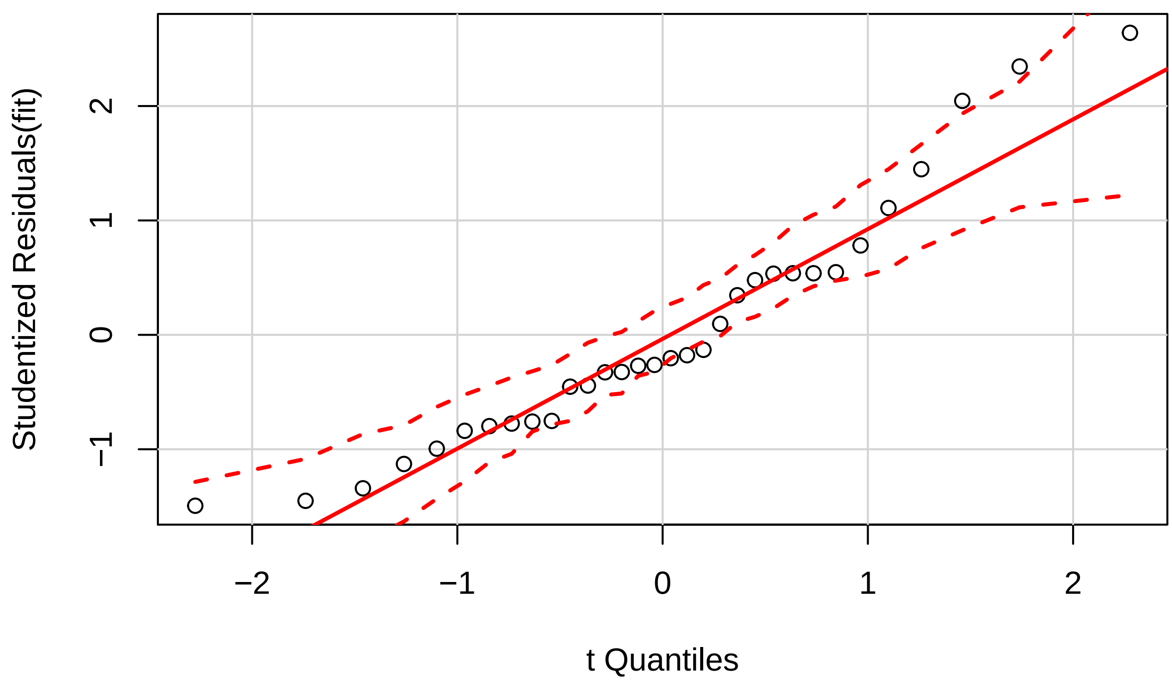



```
##                        StudRes       Hat      CookD
## Lincoln Continental -0.1310101 0.2846012 0.03764576
## Chrysler Imperial    2.0449828 0.2564037 0.50793293
## Fiat 128             2.3459187 0.1113746 0.34384709
## Toyota Corolla       2.6396909 0.1081504 0.37202500
## Toyota Corona       -1.3415216 0.2770490 0.36601464
## Maserati Bora        0.5384225 0.4661016 0.22800092
```

Creates a bubble-plot combining the display of Studentizedresiduals, hat-values, and Cook's distance (represented in the circles).

Testing fornormality

```
qqPlot(fit)
```



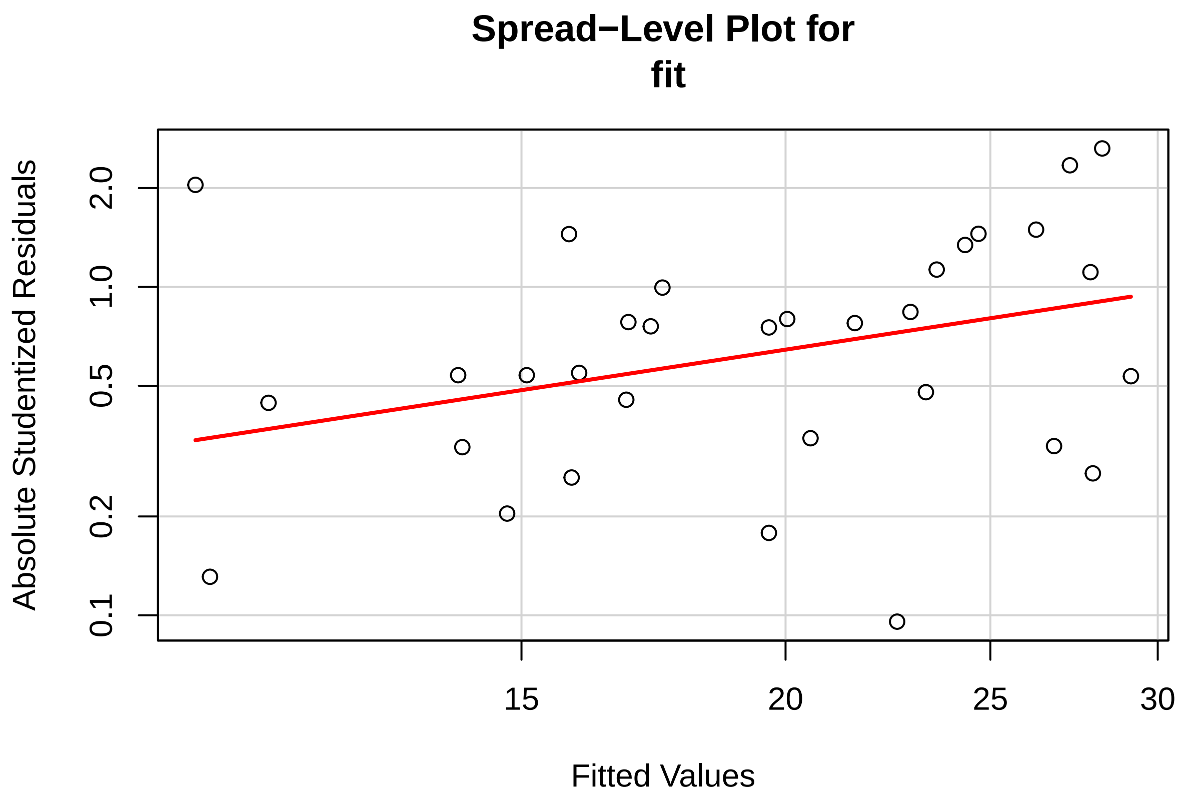
Look for the tails, points should be close to the line or within the confidence intervals. Quantileplots compare the Studentizedresiduals vsa t-distribution Other tests:shapiro.test(), mshapiro.test() in library(mvnormtest)-library(ts)

Testing for Heteroskedasticity

```
# Evaluate homoscedasticity
# non-constant error variance test
ncvTest(fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 3.407971    Df = 1     p = 0.06488218
```

```
# plot studentized residuals vs. fitted values
 spreadLevelPlot(fit)
```




```
##
## Suggested power transformation:  0.01311591
```

Breush/Pagan and Cook/Weisberg score test for non-constant error variance. Null is constant variance See also residualPlots(fit).

Testing for multicolinearity

```
vif(fit)
```

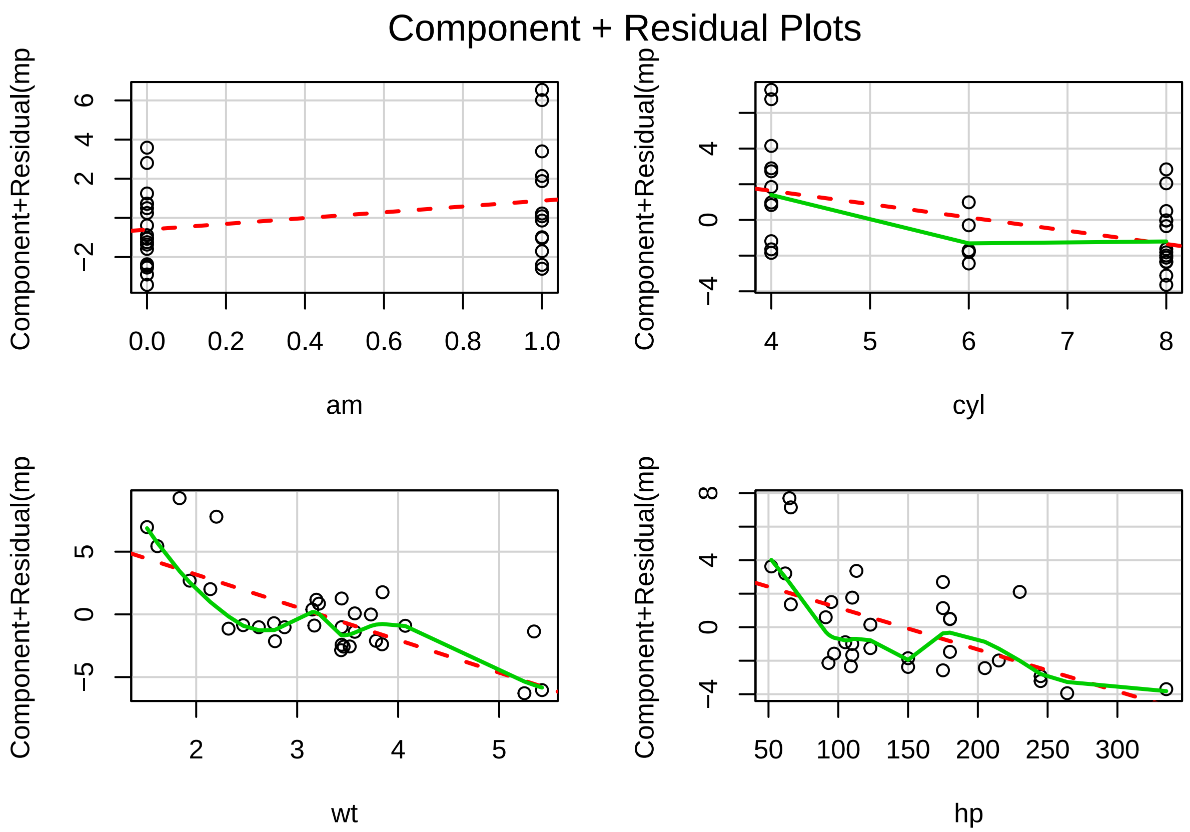
```
##       am      cyl       wt       hp
## 2.546159 5.333685 3.988305 4.310029
```

A gvif> 4 suggests collinearity. "When there are strong linear relationships among the predictors in a regression analysis, the precision of the estimated regression coefficients in linear models declines compared to what it would have been were the predictors uncorrelated with each other" (Fox:359)

Evaluate Nonlinearity

```r
# component + residual plot
crPlots(fit)
```

```
## Warning in smoother(.x, partial.res[, var], col = col.lines[2], log.x =
## FALSE, : could not fit smooth
```




```r
# Ceres plots
# ceresPlots(fit)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: mpg
##           Df Sum Sq Mean Sq F value    Pr(>F)
## am         1 405.15  405.15 64.3483 1.277e-08 ***
## cyl        1 449.53  449.53 71.3976 4.619e-09 ***
## wt         1  80.32   80.32 12.7561  0.001358 **
## hp         1  21.05   21.05  3.3432  0.078553 .
## Residuals 27 170.00    6.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test for Autocorrelated Errors

```r
durbinWatsonTest(fit)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.1412969       1.61503   0.122
##  Alternative hypothesis: rho != 0
```

Global test of model assumptions

The gvlma( ) function in the gvlma package, performs a global validation of linear model assumptions as well separate evaluations of skewness, kurtosis, and heteroscedasticity.

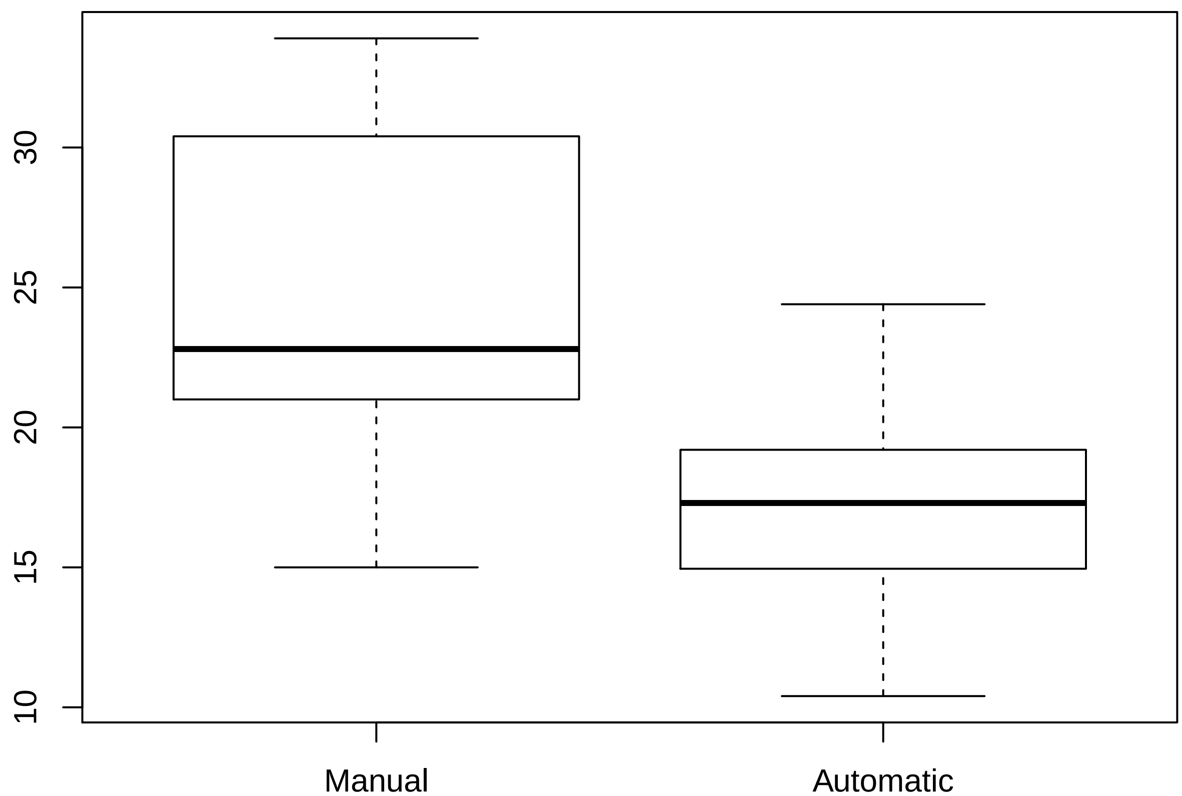
```r
library(gvlma)
gvmodel <- gvlma(fit)
summary(gvmodel)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4765 -1.8471 -0.5544  1.2758  5.6608
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.14654    3.10478  11.642 4.94e-12 ***
## am           1.47805    1.44115   1.026   0.3142
## cyl         -0.74516    0.58279  -1.279   0.2119
## wt          -2.60648    0.91984  -2.834   0.0086 **
## hp          -0.02495    0.01365  -1.828   0.0786 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.509 on 27 degrees of freedom
## Multiple R-squared:  0.849,  Adjusted R-squared:  0.8267
## F-statistic: 37.96 on 4 and 27 DF,  p-value: 1.025e-10
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = fit)
##
##                         Value  p-value                   Decision
## Global Stat        1.174e+01 0.019384 Assumptions NOT satisfied!
## Skewness           3.052e+00 0.080644    Assumptions acceptable.
```

```
## Kurtosis           7.124e-04 0.978707    Assumptions acceptable.
## Link Function      8.366e+00 0.003823 Assumptions NOT satisfied!
## Heteroscedasticity 3.225e-01 0.570086    Assumptions acceptable.
```

Side-by-side box plots

```
mtcars_vars <- mtcars[, c(1, 6, 7, 9)]
mar.orig <- par()$mar  # save the original values
par(mar = c(2, 2, 2, 2))  # set your new values
boxplot(mtcars_vars[mtcars_vars$am == 1, ]$mpg, mtcars_vars[mtcars_vars$am ==
    0, ]$mpg, names = c("Manual", "Automatic"))
```



Context

You work for Motor Trend, a magazine about the automobile industry. Looking at a data set of a collection of mtcars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

."Is an automatic or manual transmission better for MPG"

."Quantify the MPG difference between automatic and manual transmissions"

Question

Take the mtcars data set and write up an analysis to answer their question using regression models and exploratory data analyses.

Your report must be:

.Written as a PDF printout of a compiled (using knitr) R markdown document.

.Brief. Roughly the equivalent of 2 pages or less for the main text. Supporting figures in an appendix can be included up to 5 total pages including the 2 for the main report. The appendix can only include figures.

.Include a first paragraph executive summary.

Upload your PDF by clicking the Upload button below the text box.

Peer Grading

Did the student interpret the coefficients correctly? Did the student do some exploratory data analyses? Did the student fit multiple models and detail their strategy for model selection? Did the student answer the questions of interest or detail why the question(s) is (are) not answerable? Did the student do a residual plot and some diagnostics? Did the student quantify the uncertainty in their conclusions and/or perform an inference correctly? Was the report brief (about 2 pages long) for the main body of the report and no longer than 5 with supporting appendix of figures? Did the report include an executive summary? Was the report done in Rmd (knitr)?