Motor Trend magazine - Data analysis of influence on MPG for Automatic vs. Manual Transmission.

by jmvilaverde

Thursday, June 18, 2015

Executive summary (First paragraph)

For all models evaluated that have a P-value < 0.05 Manual transmission is better for MPG…

1.Initial Exploratory Data Analysis

Structure from ?mtcars and values for factors

Format: *A data frame with 32 observations on 11 variables.*

Variables	Units	Values
mpg	**Miles/(US) gallon**	
cyl	Number of cylinders (4,6,8)	4, 6, 8
disp	Displacement (cu.in.)	
hp	Gross horsepower	
drat	Rear axle ratio	
wt	Weight (lb/1000)	
qsec	1/4 mile time	
vs	V/S -> V motor or straight motor	0, 1
am	**Transmission (0 = automatic, 1 = manual)**	0, 1
gear	Number of forward gears	3, 4, 5
carb	Number of carburetors	1, 2, 3, 4, 6, 8

Correlation between mpg and am is 0.5998324. (Closer to -1 or 1 is stronger relationship, when is 0 implies no linear relationship).

2.Model proposal and analysis

First comparation, model with one predictor vs. multivariable Linear regression model formula:

One predictor	$Y_i = \beta_0 + \beta_1 X_i$	Model Initial
Multivariable	$Y_i = \Sigma_{k=1}^{p} X_{ik}\beta_j + \epsilon_i$	Model Complete

In multivariable regression analysis you must evaluate the consecuences to throwing variables that aren't related to the outcome and consecuences to omitting variables that are related to the outcome.

Analysis of Model Initial: $mpg_i = \beta_0 + \beta_{am} am_i$

View Figure C1.Summary Detail Model Initial.

```
##
## 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
```

- Intercept and coefficients estimated: 17.1473684, 7.2449393
- P-value intercept and coefficients: 1.13398345198884e-15, 0.000285020743935068 < 0.05 are good p-value for the model.
- p-value Model: 0.000285020743935069. < 0.05 is a good p-value for the model.
- R^2: 0.3597989. **This low value indicates that model Initial is not a good fit for the data.**

Analysis of model Complete: $mpg = \beta_{cyl}cyl + \beta_{dips}disp + \beta_{hp}hp + \beta_{drat}drat + \beta_{wt}wt + \beta_{qsec}qsec + \beta_{vs}vs + \beta_{am}am + \beta_{gear}gear + \beta_{carb}car$

View Figure C1.Summary Detail Model Initial.

```
##
## 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
```

- Intercept and coefficients estimated: 12.3033742, -0.1114405, 0.0133352, -0.0214821, 0.787111, -3.7153039, 0.8210407, 0.3177628, 2.5202269, 0.655413, -0.1994193
- P-value intercept and coefficients: 0.518124396898475, 0.916087375515962, 0.463488650353868, 0.334955314116978, 0.6352778979695, 0.0632521511445564, 0.273941269972363, 0.881423471976984, 0.233989710706796, 0.665206434293021, 0.812178712952693 > 0.05 are bad p-value for the model.
- p-value Model: 5.03444973840481e-10. < 0.05 is a good p-value for the model.
- R^2: 0.8690158. This high value indicates that the model Complete is a good fit to the data.

The p-value for intercept and coefficients indicates that model Complete is not a good model for the data.

Find a better model using step function:

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

- step(lm(mpg ~ ., data=mtcars)): lm, mpg ~ wt + qsec + am, mtcars

Step model: $mpg = \beta_{prop.wt} wt + \beta_{prop.qsec} qsec + \beta_{prop.am} am$

- Intercept and coefficients estimated: 9.6177805, -3.9165037, 1.225886, 2.9358372
- P-value intercept and coefficients: 0.177915165458584, 6.95271111117156e-06, 0.000216173705201939, 0.0467155099194557 > 0.05 are bad p-value for the model.
- p-value Model: 2.03846775453476e-12. < 0.05 is a good p-value for the model.
- R^2: 0.8496636. This high value indicates that the model Step is a good fit to the data.

Multivariable linear model formula:

Model Complete: $mpg = \beta_{cyl}cyl + \beta_{dips}disp + \beta_{hp}hp + \beta_{drat}drat + \beta_{wt}wt + \beta_{qsec}qsec + \beta_{vs}vs + \beta_{am}am + \beta_{gear}gear + \beta_{carb}carb$

For complete model, it is the only that has a significative effect.

> With this linear model we have a P-value over 0.05 on Intercept and all the coefficents, and under 0.05 for overall model. R^2 is 0.869 for the model, this high value indicates that the model is a good fit to the data.

> For linear regression model *Complete*, the expected difference in MPG is `r` when the car have manual transmission in comparation to the same car with automatic transmission.

```
#Confidence Interval
sumCoef.Complete <- summary(model.Complete)$coef
confInterval.Model.Complete <- sumCoef.Complete[9,1] + c(-1,1) * qt(0.975, df=model.Complete$df) * sumCo
```

Do a comparation between the 3 models

```
library(car)
anova(model.Initial, model.Step, model.Complete)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 39.2687 8.025e-08 ***
## 3     21 147.49  7     21.79  0.4432    0.8636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

View Figure 2 for plot with regression line.

3.Basic regression model with additive Gaussian errors. Into point 2 is obtained a linear regression model that fits the data, but doesn't take in consideration the impact of the others variables. We are going to analyze gaussian errors for Initial Model.

Selected variables:

- **Predictor**: X = am, Transmission with values 0 for automatic, 1 for manual.
- **Outcome**: Y = mpg, Miles/(US) gallon

Probabilistic model for linear regression:

- $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ -> $mpg_i = \beta_0 + \beta_1 am_i + \epsilon_i$
- ϵ_i are assumed iid $N(\mu_i, \sigma^2)$.
- Note, $E[Y_i|X_i = x_i] = \mu_i = \beta_0 + \beta_1 x_i$ and $Var(Y_i|X_i = x_i) = \sigma^2$.
- $\hat{\beta}_0 = \bar{Y} + \hat{\beta}_1 \bar{X}$ and $\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}$.

Residuals analysis:

- Observed outcome i is Y_i at a predictor value X_i.
- Predicted outcome i is $\hat{Y}_i$ at a predictor value X_i is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
- Residual is the difference between observed an predicted: $e_i = Y_i - \hat{Y}$, the vertical distance between the observed data point and the regression line.
- Least squares minimizes $\Sigma_{i=1}^n e_i^2$.
- e_i can be thought of as estimates of the ϵ_i.

```
#Calculate residuals
e.ModelInitial <- resid(model.Initial)

y <- mtcars$mpg
#Calculate predicted y (mpg)
yhat <- predict(model.Initial)
#Calculate max difference between residual and observed Y - predicted Y (y hat)
max(abs(e.ModelInitial -(y - yhat)))
```

```
## [1] 4.840572e-14
```

```
#Calculate residuals
e.ModelComplete <- resid(model.Complete)

#Calculate predicted y (mpg)
yhat <- predict(model.Complete)
#Calculate max difference between residual and observed Y - predicted Y (y hat)
max(abs(e.ModelComplete -(y - yhat)))
```

```
## [1] 6.57252e-14
```

```
#Calculate residuals
e.Model.Step <- resid(model.Step)

#Calculate predicted y (mpg)
yhat <- predict(model.Step)
#Calculate max difference between residual and observed Y - predicted Y (y hat)
max(abs(e.Model.Step -(y - yhat)))
```

```
## [1] 6.439294e-14
```

Figure 3 Plot of residuals:

Formula to estimate residual variation:

- ML estimate of σ^2 is $\frac{1}{n}\Sigma_{i=1}^n e_i^2$
- For $E[\hat{\sigma}^2] = \sigma^2$ most people use $\frac{1}{n-2}\Sigma_{i=1}^n e_i^2$

```
#Calculate variation
n <- length(y)
var.e.Model.Initial <- sqrt((1/(n-2))*sum((e.ModelInitial^2)))
var.e.Model.Initial
```

```
## [1] 4.902029
```

```r
#R function to calculate residual variation
summary(model.Initial)$sigma
```

```
## [1] 4.902029
```

Model Initial has a residual variation of 4.9020288.

Total variation. Formula Total variation = Residual variation + Regression Variation:

- $\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2 = \Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$

Define the percent of total variation described by the model as:

- $R^2 = \frac{\Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\Sigma_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}$

Relation between R^2 and r (the correlation):

Recall that: $(\hat{Y}_i - \bar{Y}) = \hat{\beta}_1(X_i - \bar{X})$ so that

- $R^2 = \frac{\Sigma_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2} = \hat{\beta}_1^2 \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})}{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2} = Cor(Y, X)^2$

```r
#Calculate R2
R2.Model.Initial <- sum((yhat - mean(y))^2)/sum((y-mean(y))^2)
R2.Model.Initial
```

```
## [1] 0.8496636
```

Inference in regression

Create confidence intervals and perform hypothesis tests.

```r
#Calculation of coefficients
# sigma <- var.e.Model.Initial
# ssx <- sum((x - mean(x))^2)
# seBeta0 <- (1/n + mean(x)^2/ssx) ^ 0.5 * sigma
# seBeta1 <- sigma / sqrt(ssx)
# tBeta0 <- beta0 / seBeta0
# tBeta1 <- beta1 / seBeta1
# pBeta0 <- 2 * pt(abs(tBeta0), df=n-2, lower.tail=FALSE)
# pBeta1 <- 2 * pt(abs(tBeta1), df=n-2, lower.tail=FALSE)
# coefTable <- rbind(c(beta0, seBeta0, tBeta0, pBeta0), c(beta1, seBeta1, tBeta1, pBeta1))
# colnames(coefTable) <- c("Estimate","Std.Error", "t value", "P(>|t|)")
# rownames(coefTable) <- c("(Intercept)", "x")
# coefTable
```

Final Analysis

Is an automatic or manual transmission better for MPG?

For model Initial, model Complete and model Step:

> The manual transmission is better for MPG.

Quantify the MPG difference between automatic and manual transmissions.

Model Initial:

```
sumCoef <- summary(model.Initial)$coef
confInterval.Model.Initial <- sumCoef[1,1] + c(-1,1) * qt(0.975, df=model.Initial$df) * sumCoef[1,2]
```

> With 95% confidence, we estimate that a manual transmission results in a 14.8506236 to 19.4441132 increase in MPG comparing to use of automatic transmission for the Model Initial.

Model Complete:

```
sumCoef.Complete <- summary(model.Complete)$coef
confInterval.Model.Complete <- sumCoef.Complete[1,1] + c(-1,1) * qt(0.975, df=model.Complete$df) * sumC
```

> With 95% confidence, we estimate that a manual transmission results in a -26.6225974 to 51.2293458 increase in MPG comparing to use of automatic transmission for the Model Complete.

Model Proposed:

```
sumCoef.Step <- summary(model.Step)$coef
confInterval.Model.Step <- sumCoef.Step[1,1] + c(-1,1) * qt(0.975, df=model.Step$df) * sumCoef.Step[1,2
```

> With 95% confidence, we estimate that a manual transmission results in a -4.6382995 to 23.8738605 increase in MPG comparing to use of automatic transmission for the Model Step.

Main Body + Apendix only figures (not more than 5)

C1.Figure Summary Model Initial.

```
summary(model.Initial)
```

```
##
## 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
```

C2.Figure Summary Model Complete.

```r
summary(model.Complete)
```

```
##
## 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
```

C3.Figure Summary Model Step.

```r
summary(model.Step)
```

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

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## am            2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Figure :

```
pairs(mtcars, panel = panel.smooth, main = "mtcars data", col=3+(mtcars$am>0))
```

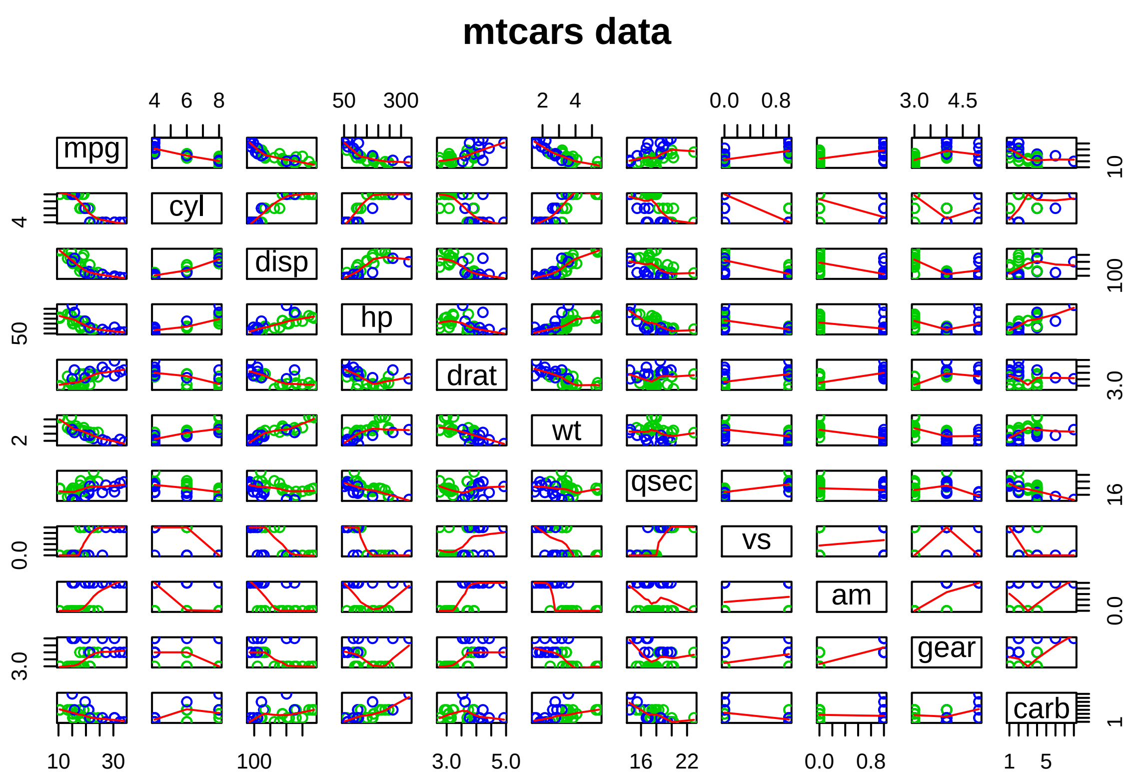
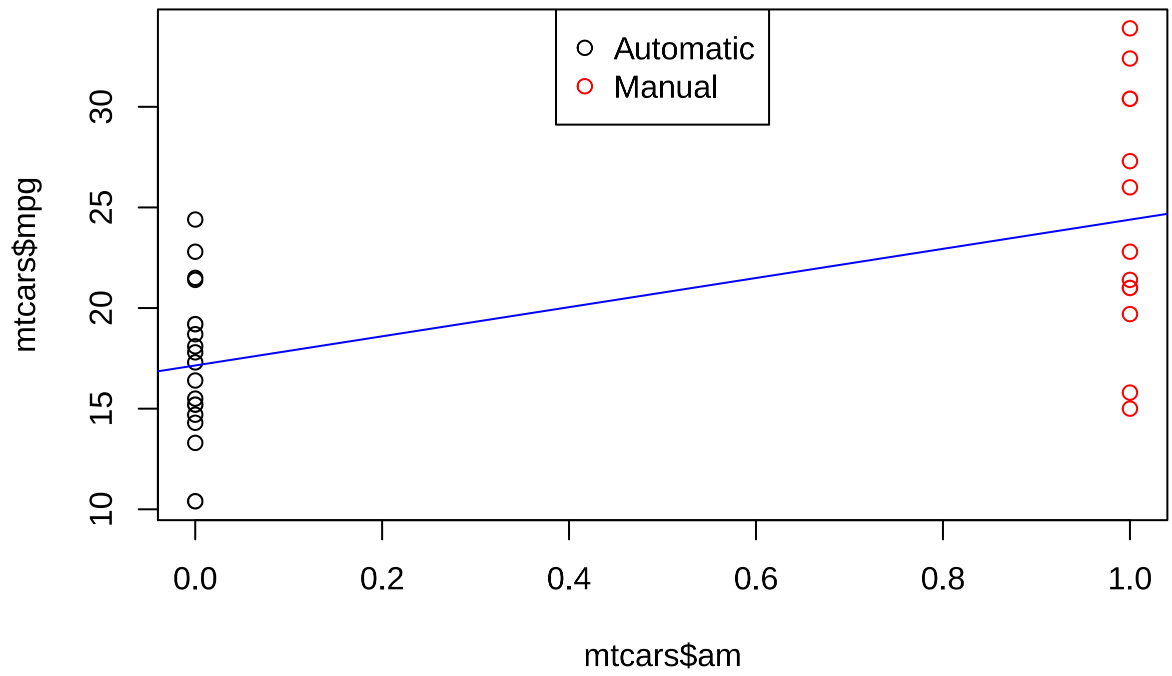
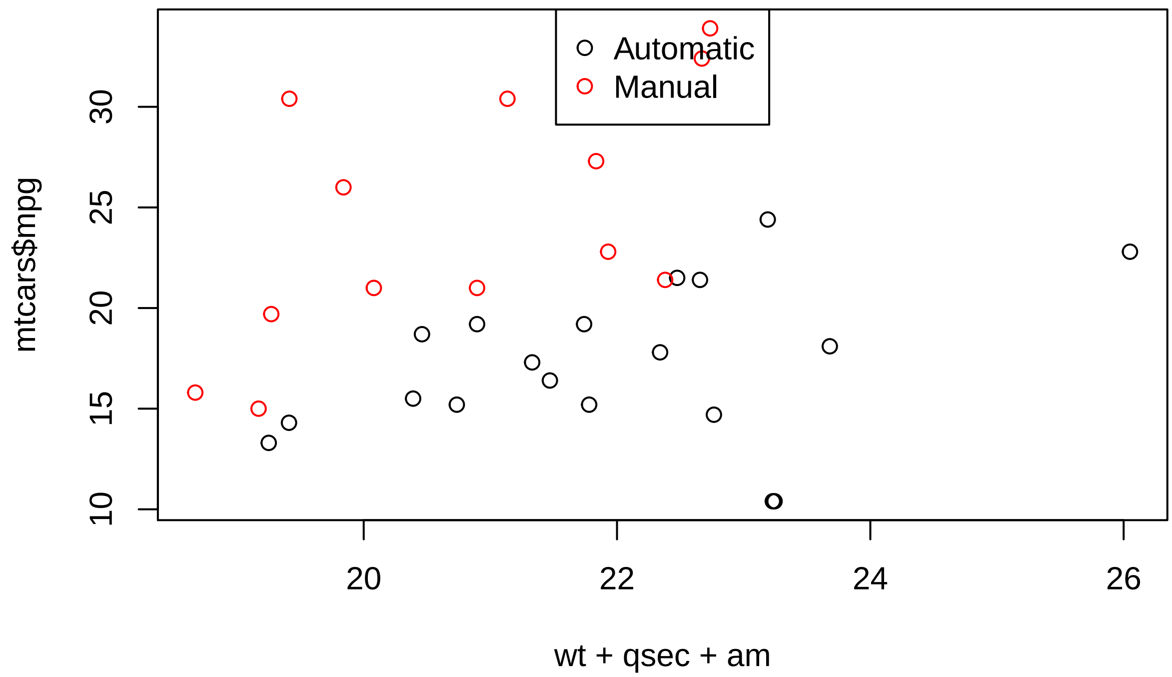


Figure 2:

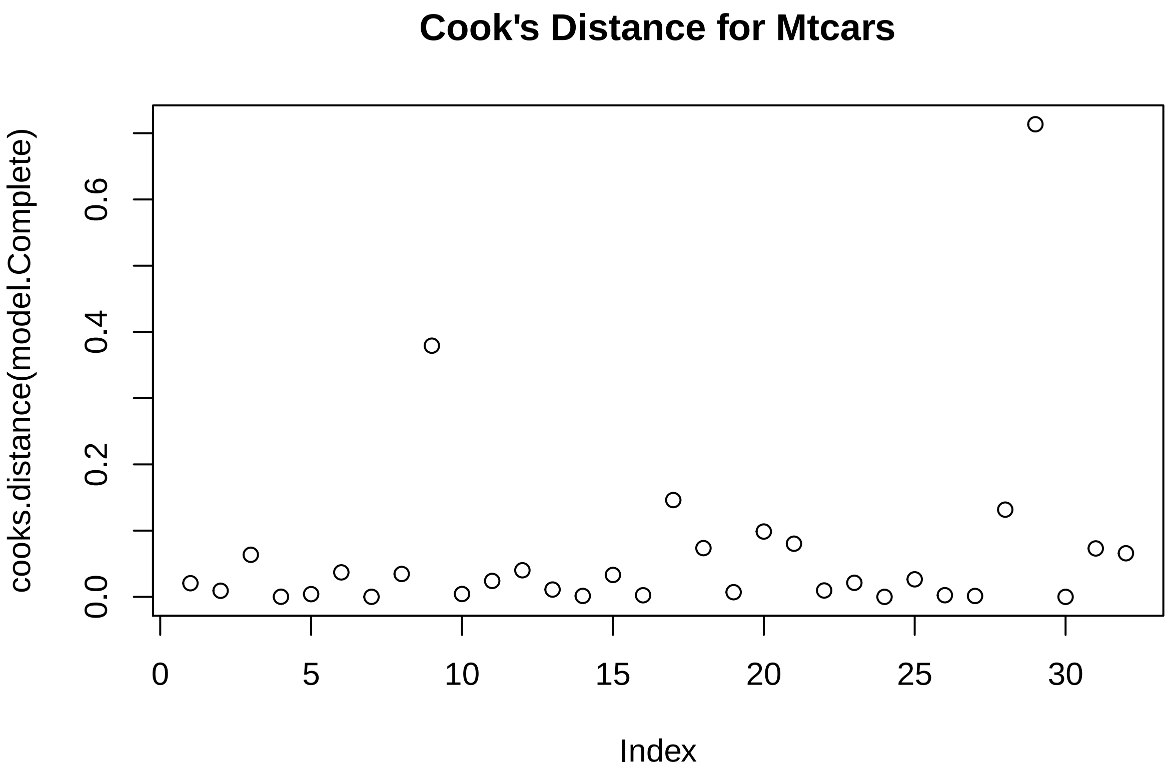
```
plot(x=mtcars$am, y=mtcars$mpg,col=mtcars$am+1)
legend("top",c("Automatic","Manual"),col=c(1,2),pch=1)
abline(model.Initial, col="blue")
```



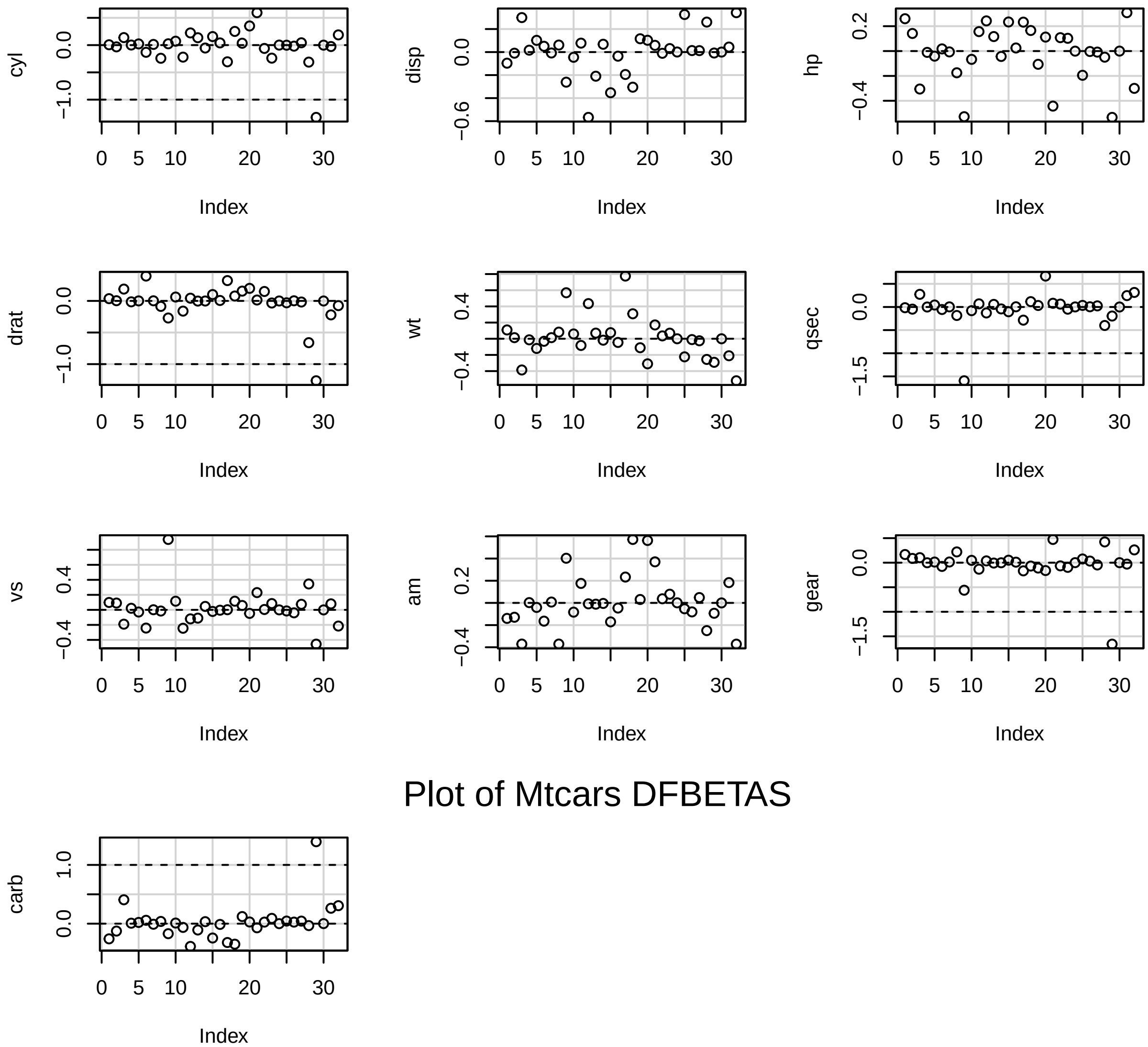

```r
with(mtcars, plot(x=wt + qsec + am, y=mtcars$mpg,col=mtcars$am+1))
legend("top",c("Automatic","Manual"),col=c(1,2),pch=1)
```



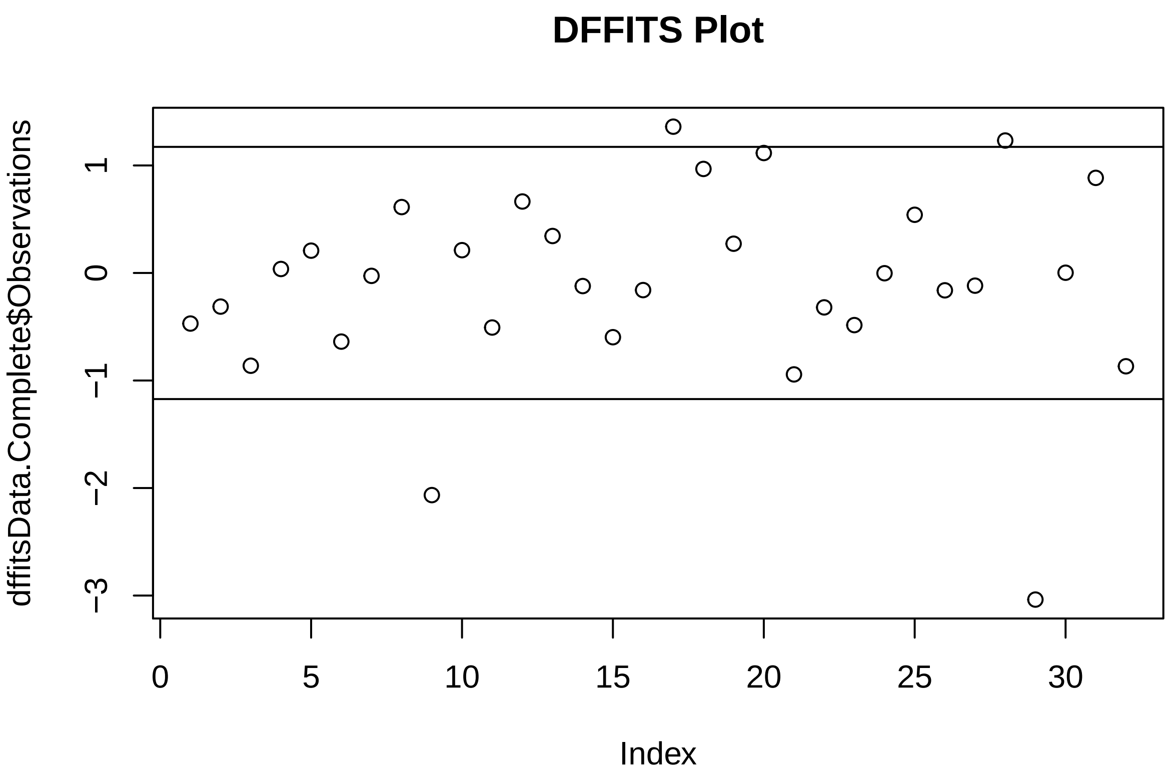

```
plot(cooks.distance(model.Complete), main="Cook's Distance for Mtcars")
```




```
dfbetasPlots(model.Complete, main="Plot of Mtcars DFBETAS")
```


cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
Index
Plot of Mtcars DFBETAS

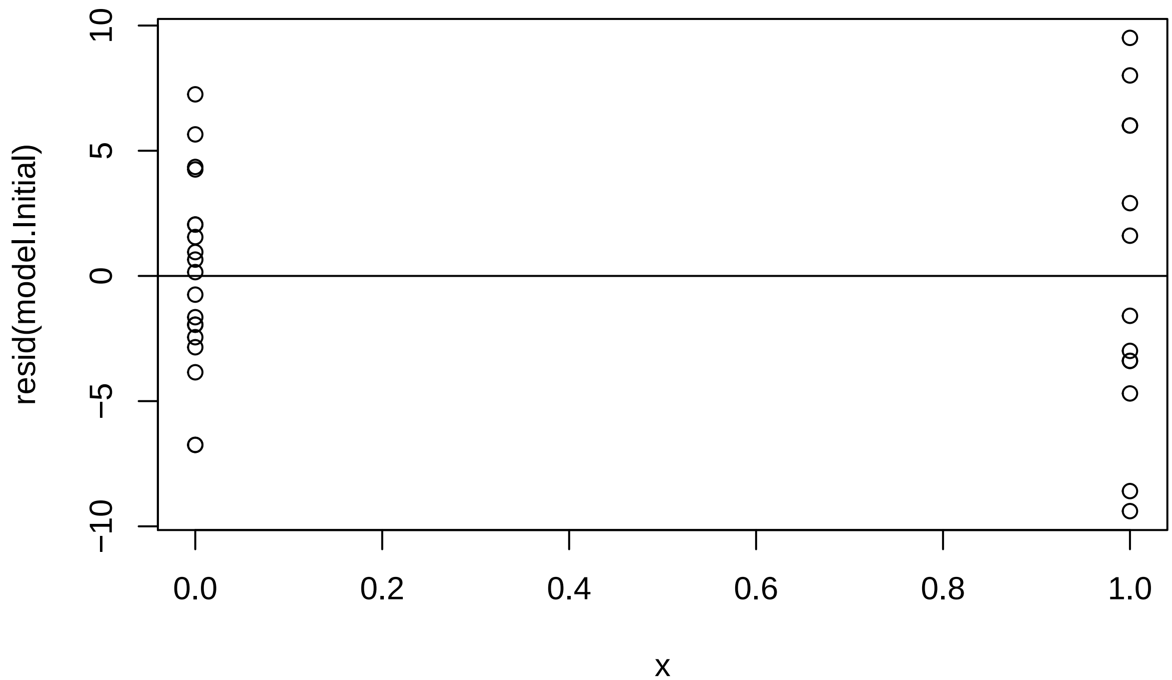
```
#Dffits
dffitsData.Complete <- as.data.frame(dffits(model.Complete))
names(dffitsData.Complete) <- c("Observations")
cutoff <- 2*sqrt(11/length(mtcars$mpg))
plot(dffitsData.Complete$Observations, main="DFFITS Plot")
abline(h=cutoff)
abline(h=-cutoff)
```



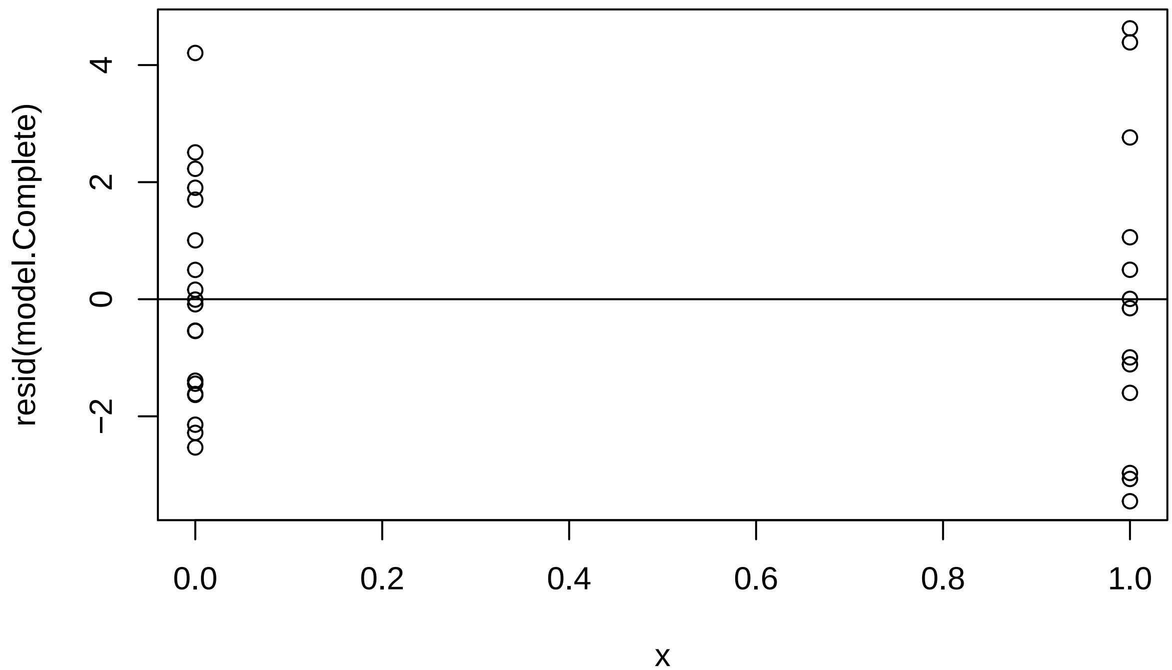

```
labels=row.names(mtcars)
```

Figure 3 Plot of residuals:

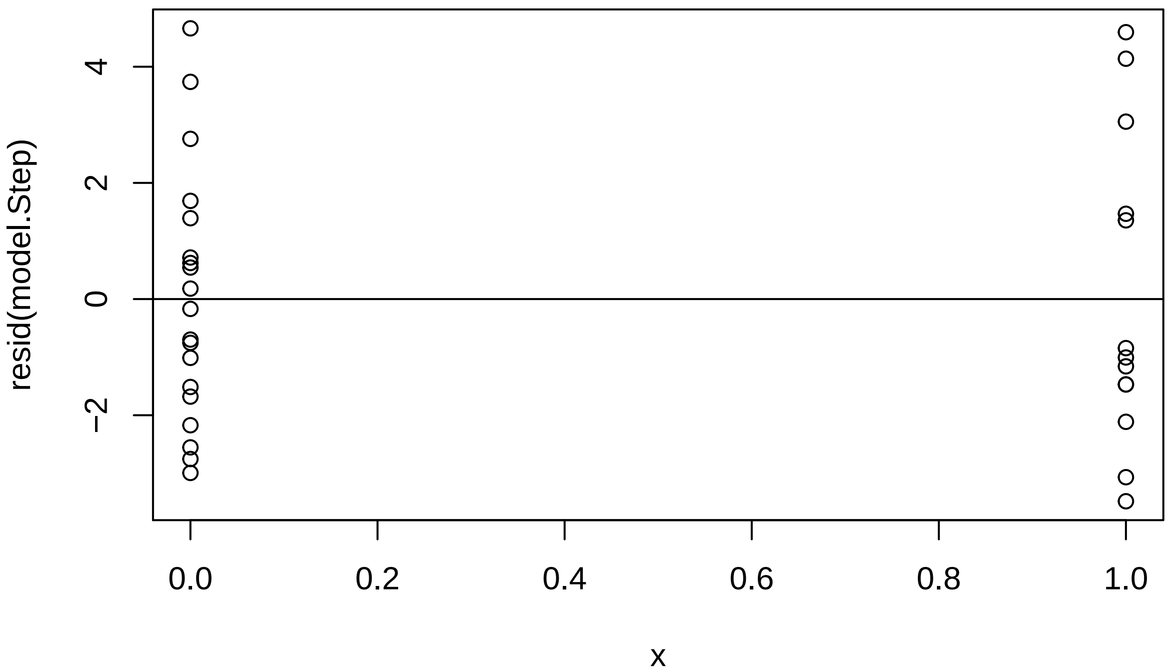
```
x <- mtcars$am
plot(x, resid(model.Initial))
abline(h=0)
```




```r
plot(x, resid(model.Complete))
abline(h=0)
```




```r
plot(x, resid(model.Step))
abline(h=0)
```


resid(model.Step)
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0
x