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

by jmvilaverde

Thursday, June 18, 2015

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.

Evaluation/feedback on the above work

- 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)?

Report brief (about 2 pages)

Executive summary (First paragraph)

Manual transmission is better for MPG…

Data adquisition

1. Data adquisition and Initial structure analysis

```
data(mtcars)

#Show colnames of mtcars
names(mtcars)
```

1.1. Variables

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

mtcars

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

Variable	Units
mpg	**Miles/(US) gallon**
cyl	Number of cylinders
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
am	**Transmission (0 = automatic, 1 = manual)**
gear	Number of forward gears
carb	Number of carburetors

```r
#Show structure of 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 ...
```

2.Initial analysis

One predictor Selected variables:

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

Linear regression model formula: Y = $\beta_0 + \beta_1 X$ -> mpg = $\beta_0 + \beta_1 am$

Create initial model:

- Calculate β_0 (Intercept): $\hat{\beta}_0 = \bar{Y} + \hat{\beta}_1 \bar{X}$
- Calculate β_1 (Slope): $\hat{\beta}_1 = Cor(Y, X)\frac{Sd(Y)}{Sd(X)}$

Fitting the best line: $\Sigma_n^{i=1}(Y_i - (\beta_0 + \beta_1 X_i))^2$ to minimize the distance.

```
#Theoretical formula
y <- mtcars$mpg
x <- mtcars$am
beta1 <- cor(y,x)*(sd(y)/sd(x))
beta0 <- mean(y) - beta1 * mean(x)

#R function
model.Initial <- lm(mpg ~ am, data=mtcars)
coeffs <- coef(model.Initial)

#Comparation
rbind(c(beta0, beta1), coeffs)
```

```
##        (Intercept)       am
##           17.14737 7.244939
## coeffs    17.14737 7.244939
```

For linear model $Y = \beta_0 + \beta_1 X$ -> $mpg = \beta_0 + \beta_1 am$

- Intercept $\beta_0 = 17.1473684$
- Slope $\beta_1 = 7.2449393$

> For this linear regression model, the expected difference in MPG is 7.2449393 when the car have manual transmission in comparation to the same car with automatic transmission. The Intercept 17.1473684 is the expected MPG of a automatic transmission car.

Multiple predictor 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.

Multivariable linear model formula:

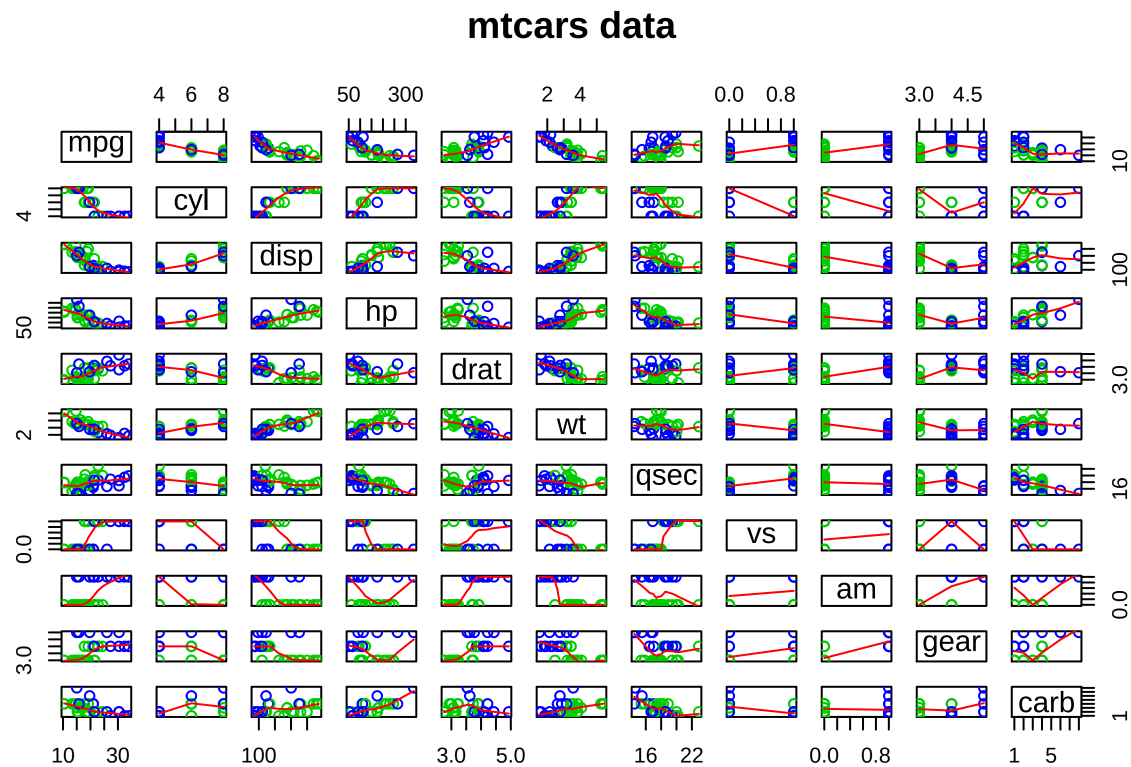
- $Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \epsilon_i = \Sigma_{k=1}^p X_{ik}\beta_j + \epsilon_i$

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

```
model.Complete <- lm(mpg ~ ., data=mtcars)
beta.am.Complete <- coef(model.Complete)["am"]
library(car)
```

> For linear regression model *Complete*, the expected difference in MPG is 2.5202269 when the car have manual transmission in comparation to the same car with automatic transmission.

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




```r
#Model Complete
model.Complete <- lm(mpg ~ ., data = mtcars)
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
```

```
#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) * sumC
```

Model to be proposed. Define a model based on Variation Inflation Factor analysis to check collinearity between variables:

```
vif.Complete <- vif(model.Complete)
vif.Complete
```

```
##       cyl      disp        hp      drat        wt      qsec        vs
## 15.373833 21.620241  9.832037  3.374620 15.164887  7.527958  4.965873
##        am      gear      carb
##  4.648487  5.357452  7.908747
```

VIF	Indication
1	No correlation
$1 < VIF < 5$	moderate correlation
$VIF > 5$	strong correlation

Remove variables with strong correlation: Use the other variables:

Proposed model: $mpg = \beta_{prop.hp}hp + \beta_{prop.drat}drat + \beta_{prop.vs}vs + \beta_{prop.am}am$

```
model.Prop <- lm(formula = mpg ~ drat + vs + am, data = mtcars)
beta.am.Prop <- coef(model.Prop)["am"]
vif.Prop <- vif(model.Prop)
sqrt(vif.Prop)
```

```
##     drat       vs       am
## 1.608598 1.144700 1.465209
```

```
summary(model.Prop)
```

```
##
## Call:
## lm(formula = mpg ~ drat + vs + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9892 -2.6090  0.2629  2.1127  6.2924
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    8.327      6.017   1.384 0.177316
## drat           1.985      1.883   1.054 0.300772
## vs             6.235      1.421   4.387 0.000148 ***
## am             4.669      1.838   2.540 0.016898 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.485 on 28 degrees of freedom
## Multiple R-squared:  0.6981, Adjusted R-squared:  0.6657
## F-statistic: 21.58 on 3 and 28 DF,  p-value: 1.922e-07
```

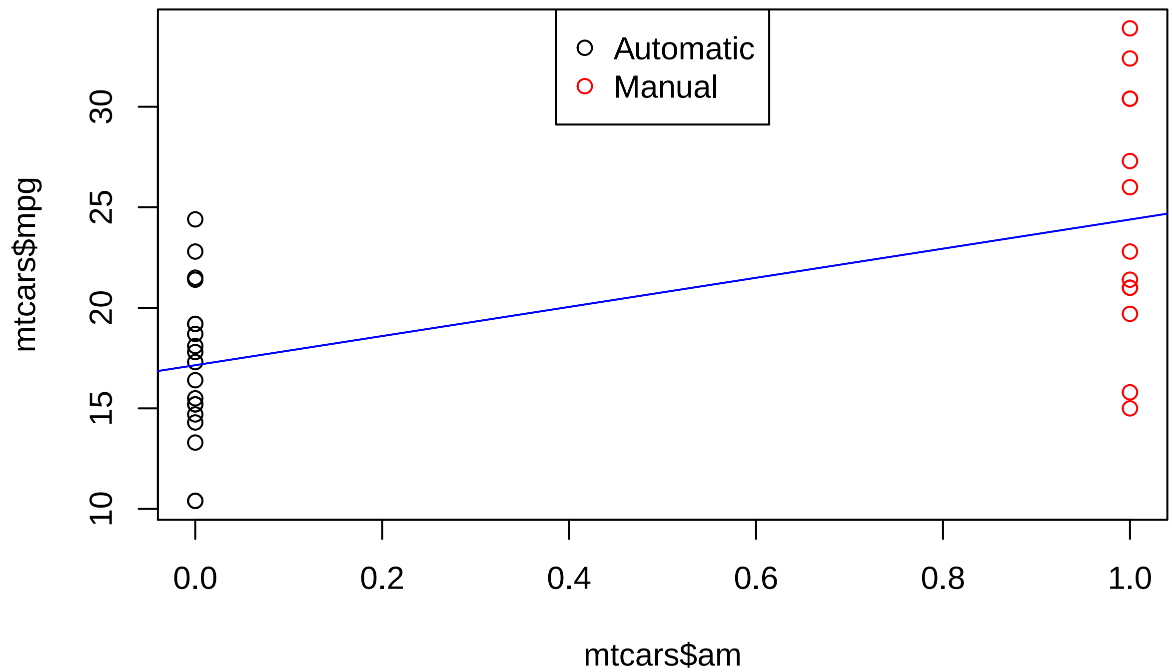
For linear regression model *Prop*, the expected difference in MPG is 4.6687251 when the car have manual transmission in comparation to the same car with automatic transmission.

Do a comparation between the 3 models

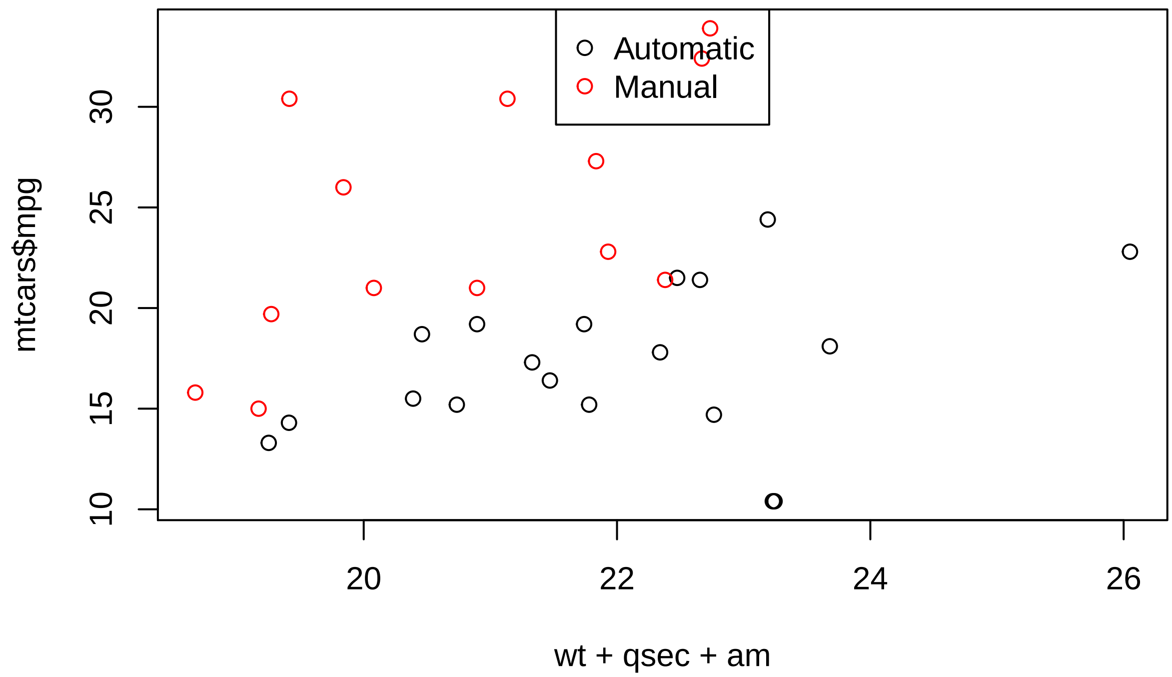
```r
library(car)
anova(model.Initial, model.Prop, model.Complete)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ drat + vs + 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     28 339.99  2    380.91 27.1164 1.517e-06 ***
## 3     21 147.49  7    192.50  3.9153  0.006983 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

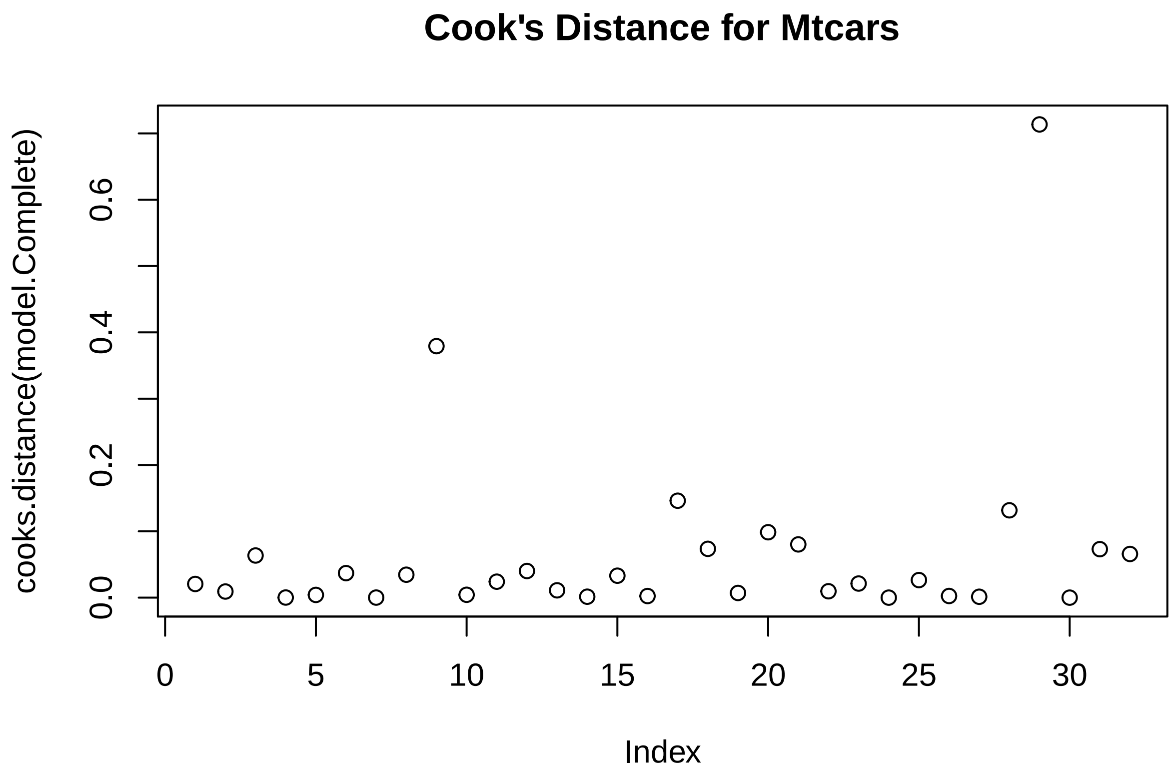
```r
plot(x=mtcars$am, y=mtcars$mpg,col=mtcars$am+1)
legend("top",c("Automatic","Manual"),col=c(1,2),pch=1)
abline(model.Initial, col="blue")
```



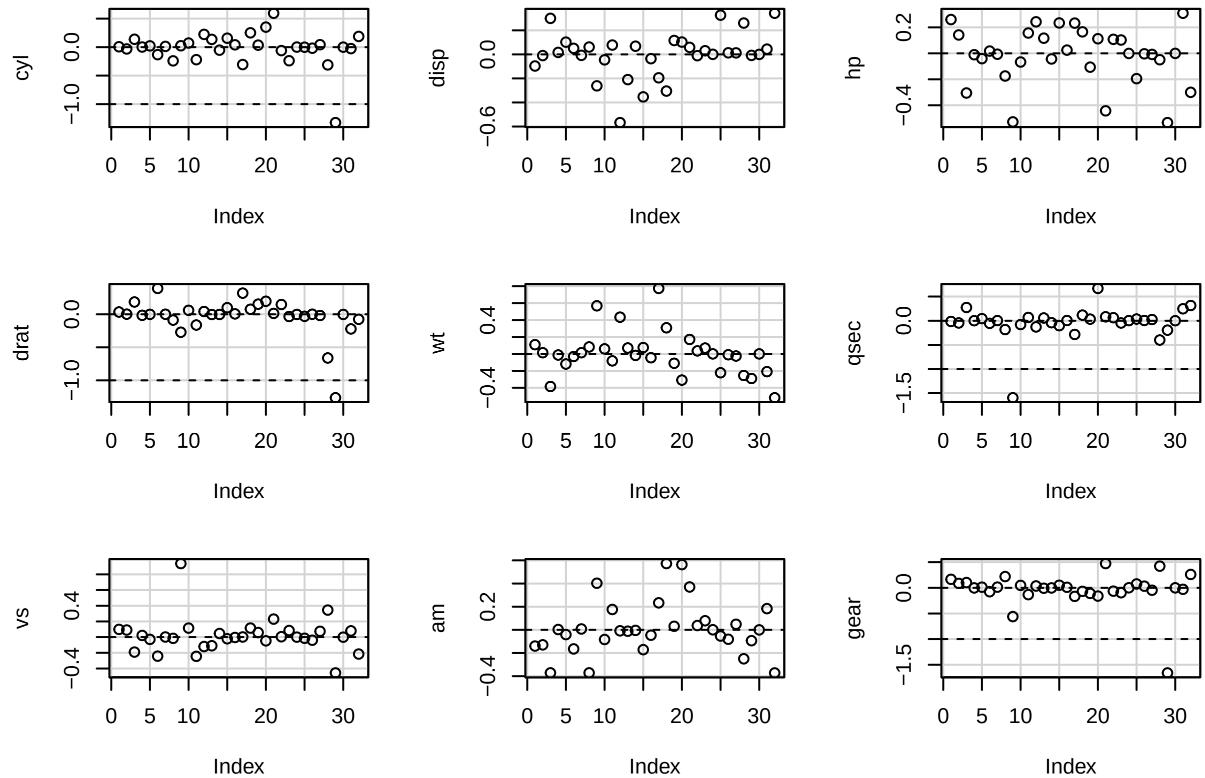

```
with(mtcars, plot(x=wt + qsec + am, y=mtcars$mpg,col=mtcars$am+1))
legend("top",c("Automatic","Manual"),col=c(1,2),pch=1)
```



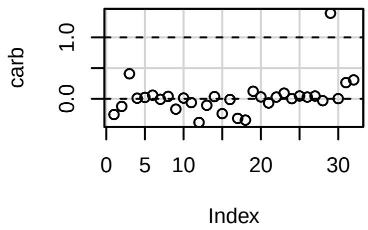

```
plot(cooks.distance(model.Complete), main="Cook's Distance for Mtcars")
```



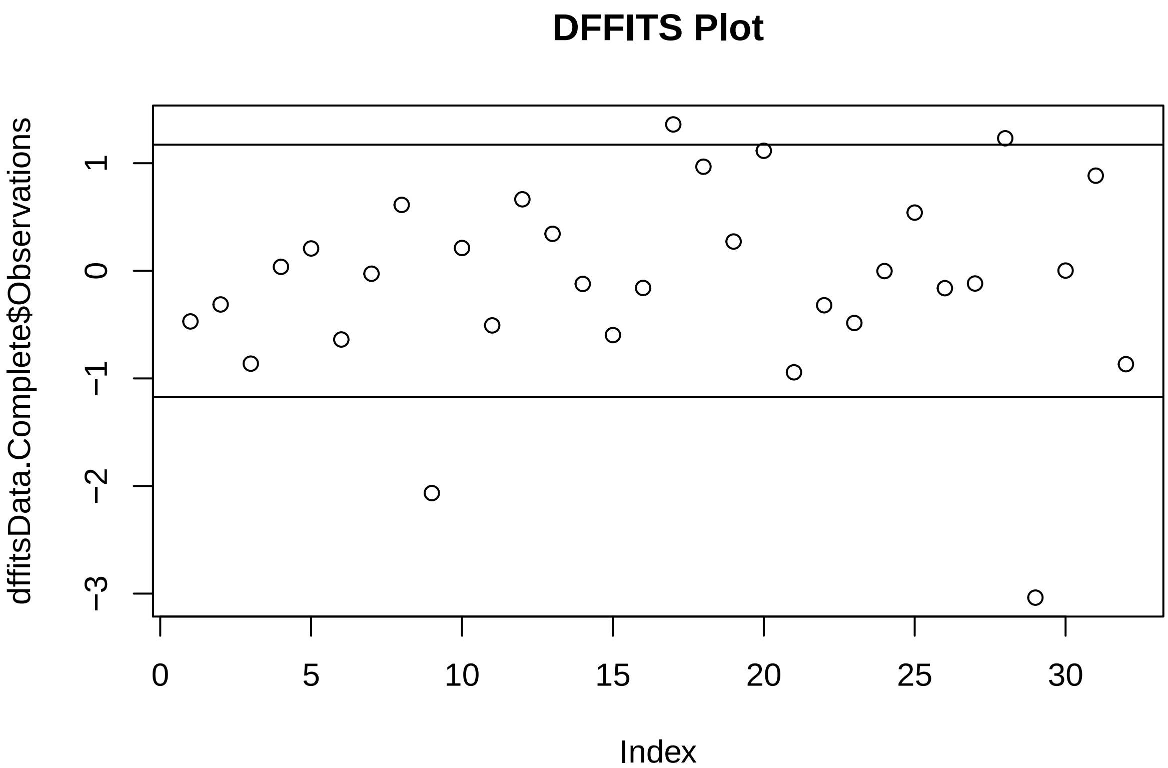

```
dfbetasPlots(model.Complete, main="Plot of Mtcars DFBETAS")
```


cyl
disp
hp
drat
wt
qsec
vs
am
gear
Index

Plot of Mtcars DFBETAS




```r
#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)
```




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

View Figure 1 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 <- y-(beta0+beta1*x)
e.ModelInitial
```

```
##  [1] -3.3923077 -3.3923077 -1.5923077  4.2526316  1.5526316  0.9526316
##  [7] -2.8473684  7.2526316  5.6526316  2.0526316  0.6526316 -0.7473684
## [13]  0.1526316 -1.9473684 -6.7473684 -6.7473684 -2.4473684  8.0076923
## [19]  6.0076923  9.5076923  4.3526316 -1.6473684 -1.9473684 -3.8473684
## [25]  2.0526316  2.9076923  1.6076923  6.0076923 -8.5923077 -4.6923077
## [31] -9.3923077 -2.9923077
```

```
#R function for residuals
resid(model.Initial)
```

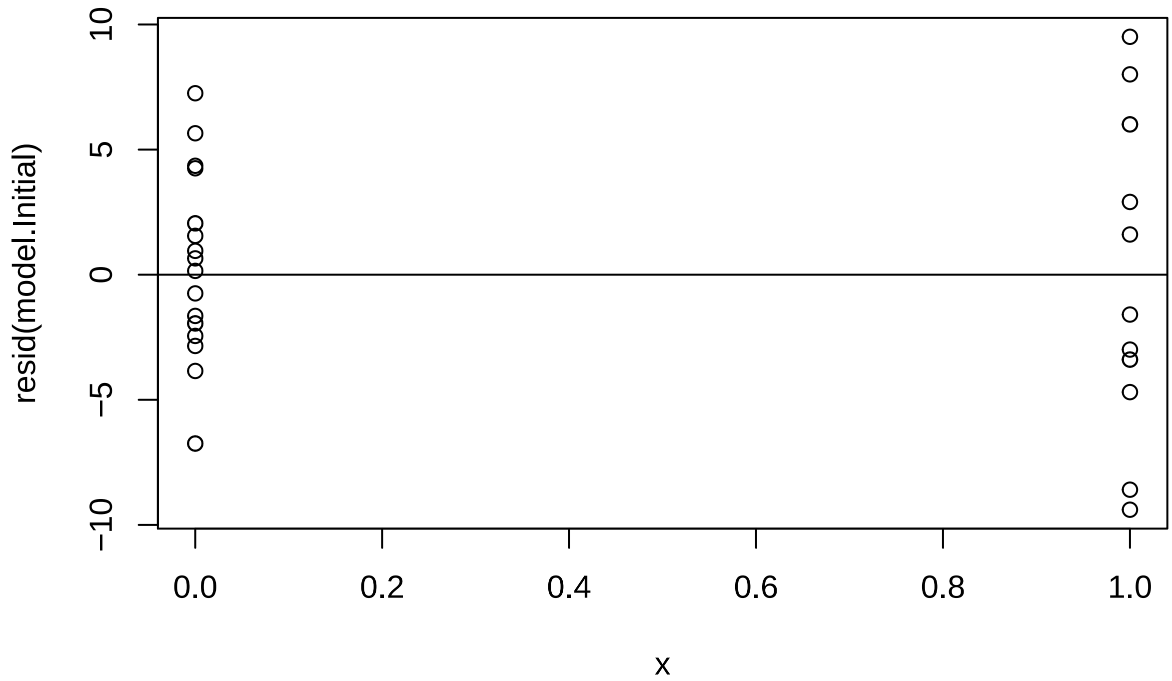
```
##           Mazda RX4       Mazda RX4 Wag          Datsun 710
##          -3.3923077          -3.3923077          -1.5923077
##      Hornet 4 Drive   Hornet Sportabout             Valiant
##           4.2526316           1.5526316           0.9526316
##          Duster 360           Merc 240D            Merc 230
##          -2.8473684           7.2526316           5.6526316
##            Merc 280           Merc 280C          Merc 450SE
##           2.0526316           0.6526316          -0.7473684
##          Merc 450SL         Merc 450SLC  Cadillac Fleetwood
##           0.1526316          -1.9473684          -6.7473684
## Lincoln Continental   Chrysler Imperial            Fiat 128
##          -6.7473684          -2.4473684           8.0076923
##         Honda Civic      Toyota Corolla       Toyota Corona
##           6.0076923           9.5076923           4.3526316
##    Dodge Challenger         AMC Javelin          Camaro Z28
##          -1.6473684          -1.9473684          -3.8473684
##    Pontiac Firebird           Fiat X1-9       Porsche 914-2
##           2.0526316           2.9076923           1.6076923
##        Lotus Europa      Ford Pantera L        Ferrari Dino
##           6.0076923          -8.5923077          -4.6923077
##       Maserati Bora          Volvo 142E
##          -9.3923077          -2.9923077
```

```
#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] 3.552714e-15
```

Plot of residuals:

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



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 $\frac{1}{n-2}\Sigma_{i=1}^{n} e_i^2$

```r
#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

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

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

```
## [1] 0.3597989
```

```
#R function for R2
cor(y,x)^2
```

```
## [1] 0.3597989
```

Inference in regression

Create confidence intervals and perform hypothesis tests.

```
#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
```

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

```
#R function for calculate coefficients
summary(model.Initial)$coef
```

```
##              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
```

Getting a confidence interval

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

```
## [1] 14.85062 19.44411
```

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

```
## [1] -26.62260  51.22935
```

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

```
## [1] -3.997996 20.651141
```

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

Final Analysis for Model Initial

This analysis doesn't take in consideration other variables that MUST be evaluated due the influence on relation between MPG and the type of transmission AM. Is required a Multivariable regression model.

Is an automatic or manual transmission better for MPG?

> The manual transmission is better for MPG.

Quantify the MPG difference between automatic and manual transmissions.

> 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.
>
> - Predictor: X = am, Transmission with values 0 for automatic, 1 for manual.
> - Outcome: Y = mpg, Miles/(US) gallon
>
> Linear regression model formula: Y = $\beta_0 + \beta_1 X$ -> mpg = $\beta_0 + \beta_1 am$

Use of other models

Conclussion for model complete

Is an automatic or manual transmission better for MPG?

> The manual transmission is better for MPG.

Quantify the MPG difference between automatic and manual transmissions.

> 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 Initial.

Conclussion for model proposed

Is an automatic or manual transmission better for MPG?

> The manual transmission is better for MPG.

Quantify the MPG difference between automatic and manual transmissions.

> With 95% confidence, we estimate that a manual transmission results in a -3.9979958 to 20.6511414 increase in MPG comparing to use of automatic transmission for the Model Initial. ***

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