library(car)

library(alr4)

# 1

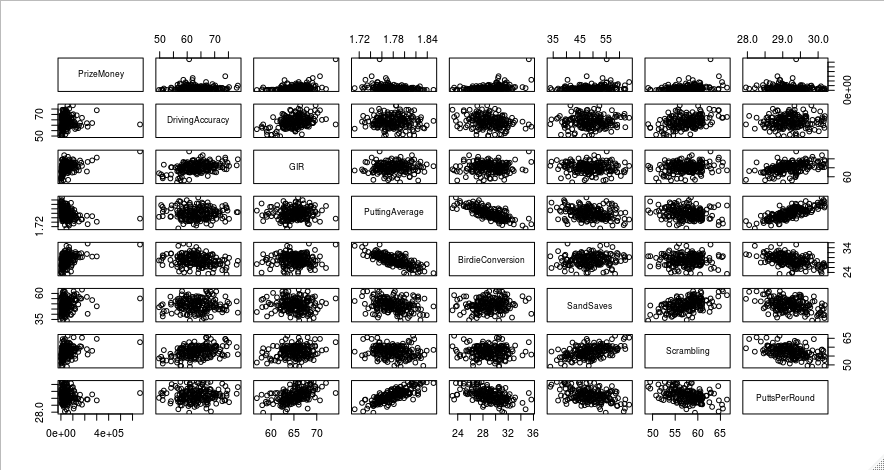
golf <- read.csv("/home/thomas/git/datascience/MSDataSci/MATH550/data/pgatour2006.csv",header=TRUE)

**# Removing unneeded columns**

golf <- subset(golf, select=-c(Name,TigerWoods,AveDrivingDistance,BounceBack))

attach(golf)

golf

pairs(PrizeMoney~., data = golf)

model\_golf <- lm(PrizeMoney ~., data=golf)

summary(model\_golf)

Call:

lm(formula = PrizeMoney ~ ., data = golf)

Residuals:

Min 1Q Median 3Q Max

-81239 -26260 -6521 17539 420230

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1165233.1 587382.9 -1.984 0.048737 \*

DrivingAccuracy -1835.8 889.2 -2.065 0.040326 \*

GIR 9671.3 3309.4 2.922 0.003899 \*\*

PuttingAverage -47435.3 521566.4 -0.091 0.927631

BirdieConversion 10426.0 3049.6 3.419 0.000771 \*\*\*

SandSaves 1182.1 744.8 1.587 0.114184

Scrambling 4741.3 2400.8 1.975 0.049749 \*

PuttsPerRound 5267.5 35765.7 0.147 0.883070

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 50140 on 188 degrees of freedom

Multiple R-squared: 0.4064, Adjusted R-squared: 0.3843

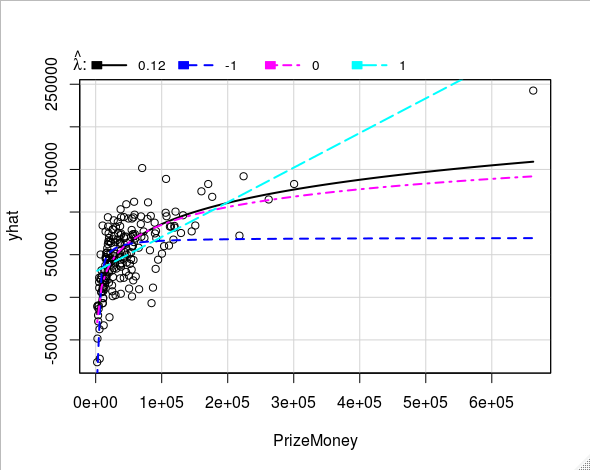
F-statistic: 18.39 on 7 and 188 DF, p-value: < 2.2e-16

**# Running a full summary of a linear model now does not give us the best R-squared value at 0.3843.**

**# Let's try a transform...**

par(mfrow = c(1,1))

invResults\_golf <- invResPlot(model\_golf)



**# The Inverse response plot picked out a visually appropriate lambda-hat of about 0.12**

**# to use as an exponent in the power transform for PrizeMoney**

**# assigning lambda-hat to a variable**

lambdaHat\_golf <- invResults\_golf$lambda[1]

p1 <- powerTransform(PrizeMoney ~.,data=golf)

**# Power transforming PrizeMoney**

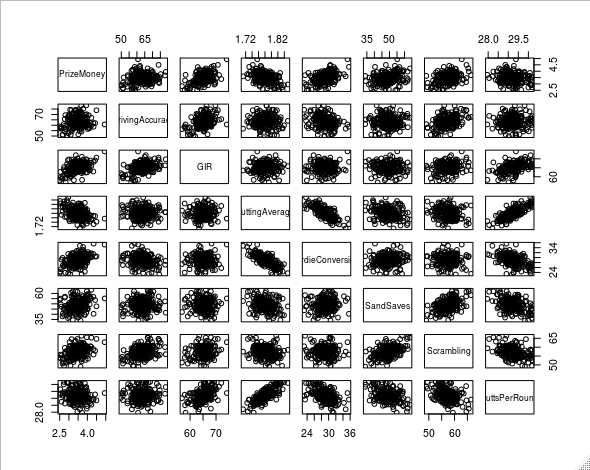
golf\_power <- golf

golf\_power$PrizeMoney <- golf\_power$PrizeMoney^lambdaHat\_golf

golf\_power

attach(golf\_power)

pairs(PrizeMoney~., data = golf\_power)



**# The power transformed pairs plot looks much better. Now, a rough linear relationship can be seen**

**# between PrizeMoney and the predictors with some multicollinearity.**

model\_golf\_power <- lm(PrizeMoney ~., data=golf\_power)

summary(model\_golf\_power)

**# The adjusted R-squared value for the new power-transformed model is a lot better at 0.5354.**

**# Now, let's try a Log Transformation on Prize Money**

**# Log transforming PrizeMoney**

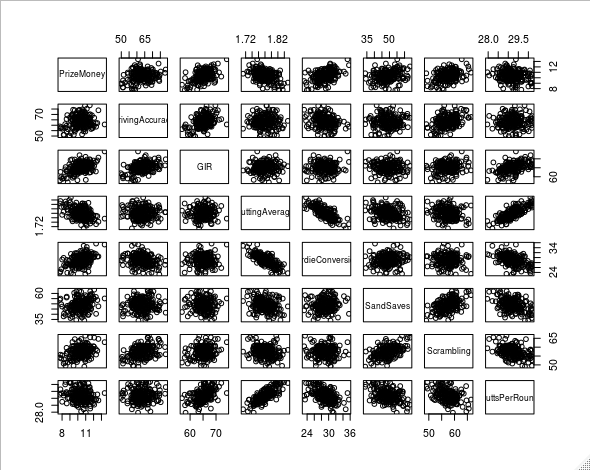
golf\_log <- golf

golf\_log$PrizeMoney <- log(golf\_log$PrizeMoney)

golf\_log

attach(golf\_log)

pairs(PrizeMoney~., data = golf\_log)



**# The log transformed pairs plot looks much better. Now, a rough linear relationship can be seen**

**# between PrizeMoney and the predictors.**

model\_golf\_log <- lm(PrizeMoney ~., data=golf\_log)

summary(model\_golf\_log)

**# The adjusted R-squared value for the new log-transformed model is 0.5412.**

**# Although the log transformation is slightly better than the power transformation given the**

**# data provided, The inverse response plot visually looks like the data would be much better**

**# explained by a power transformation. I will continue with the power transformation**

**# In the new power transformation pairs plot, there looks to be a lot of multicollinearity.**

**# Because of this, I will perform backwards step AIC and BIC to reduce the**

**# number of predictors and potential multicollinearity**

n <- length(model\_golf\_power$residuals)

backAIC <- step(model\_golf\_power,direction="backward", data=golf\_power)

Step: AIC=-503.54

PrizeMoney ~ GIR + BirdieConversion + SandSaves + Scrambling

Df Sum of Sq RSS AIC

<none> 14.268 -503.54

- SandSaves 1 0.2704 14.539 -501.86

- Scrambling 1 1.1774 15.446 -489.99

- GIR 1 5.8315 20.100 -438.37

- BirdieConversion 1 6.1694 20.438 -435.10

backBIC <- step(model\_golf\_power,direction="backward", data=golf\_power, k=log(n))

Step: AIC=-488.74

PrizeMoney ~ GIR + BirdieConversion + Scrambling

Df Sum of Sq RSS AIC

<none> 14.539 -488.74

- Scrambling 1 2.5946 17.133 -461.84

- GIR 1 5.5610 20.100 -430.54

- BirdieConversion 1 6.8541 21.393 -418.32

**# The step model results driven by AIC give us back SandSaves, Scrambling, GIR, and BirdieConversion**

**# as the best predictors.**

**# For BIC, I am given Scrambling, GIR, and BirdieConversion**

**# Since there are already so few terms, the BIC approach might not be completely necessary**

**# It might be better to evaluate the models for each before making a decision on the best**

# AIC

model\_golf\_power <- lm(PrizeMoney~SandSaves+Scrambling+GIR+BirdieConversion, data=golf\_power)

summary(model\_golf\_power)

Residuals:

Min 1Q Median 3Q Max

-0.70071 -0.20124 -0.04206 0.18819 0.87718

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.313882 0.598530 -8.878 4.94e-16 \*\*\*

SandSaves 0.007647 0.004019 1.903 0.058584 .

Scrambling 0.029633 0.007464 3.970 0.000102 \*\*\*

GIR 0.066376 0.007513 8.835 6.49e-16 \*\*\*

BirdieConversion 0.081995 0.009023 9.088 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2733 on 191 degrees of freedom

Multiple R-squared: 0.5473, Adjusted R-squared: 0.5378

F-statistic: 57.73 on 4 and 191 DF, p-value: < 2.2e-16

# BIC

model\_golf\_power <- lm(PrizeMoney~Scrambling+GIR+BirdieConversion, data=golf\_power)

summary(model\_golf\_power)

Residuals:

Min 1Q Median 3Q Max

-0.68968 -0.22136 -0.02909 0.17554 0.83980

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.261075 0.601952 -8.740 1.16e-15 \*\*\*

Scrambling 0.037209 0.006357 5.854 2.05e-08 \*\*\*

GIR 0.063272 0.007383 8.570 3.38e-15 \*\*\*

BirdieConversion 0.085048 0.008939 9.514 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2752 on 192 degrees of freedom

Multiple R-squared: 0.5387, Adjusted R-squared: 0.5315

F-statistic: 74.74 on 3 and 192 DF, p-value: < 2.2e-16

pairs(PrizeMoney~SandSaves+Scrambling+GIR+BirdieConversion, data = golf\_power)

**# While the AIC has an additional term, the R-squared is slightly higher than using BIC, yet negligible -**

**# a difference between 0.5378 (AIC) and 0.5315 (BIC)**

**# Since I do not expect the span of the addition of another variable from the AIC method over the BIC method**

**# To be much larger, I will default to the predictors picked by the back step BIC method.**

**# I do this because, since the influence of an additional predictor is very small, I do not want to run the**

**# risk of overfitting the data on that extra variable now and causing the model to falter with future data.**

par(mfrow = c(2,2))

mmp(model\_golf\_power)

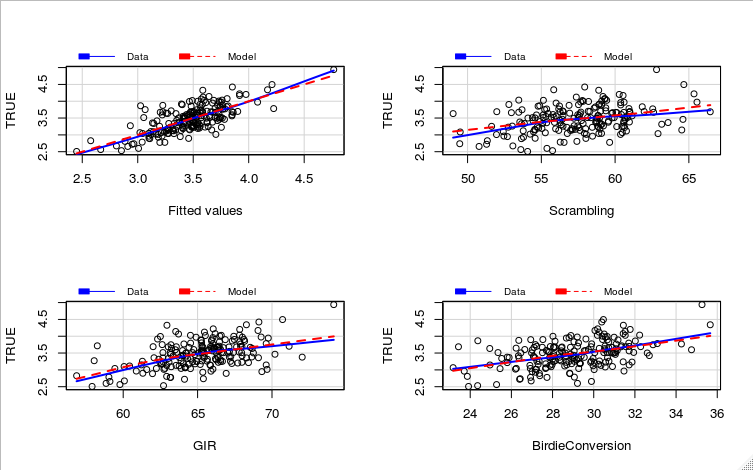
mmp(model\_golf\_power,Scrambling)

mmp(model\_golf\_power,GIR)

mmp(model\_golf\_power,BirdieConversion)

**# The Marginal Model plots for the model and the individual predictors are looking good!**

**# The model is tracking well wit the data.**



StanRes <- rstandard(model\_golf\_power)

par(mfrow = c(2,2))

plot(Scrambling,StanRes, ylab="Standardized Residuals",pch=19,col=10)

abline(h=0,lwd=2)

plot(GIR,StanRes, ylab="Standardized Residuals",pch=19,col=10)

abline(h=0,lwd=2)

plot(BirdieConversion,StanRes, ylab="Standardized Residuals",pch=19,col=10)

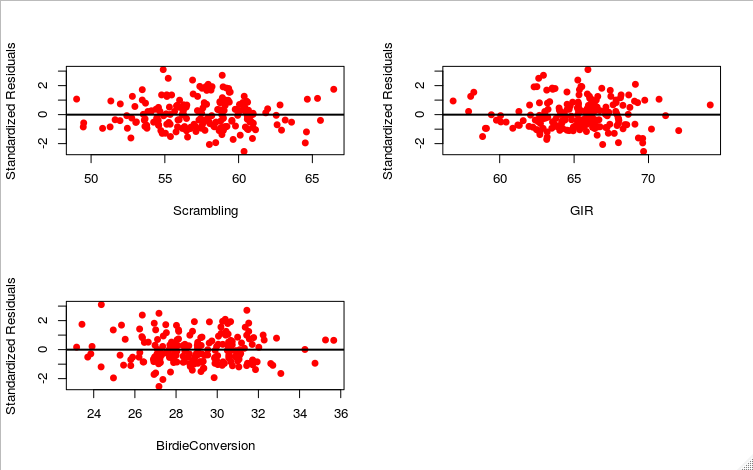
abline(h=0,lwd=2)

**# The standard residuals also are looking nice and flat.**

**# If I had to investigate the model further, I would look into Outlier Detection.**

**# There are a few points that look like they might be providing slightly more**

**# leverage than others and may need to be addressed.**



# 2

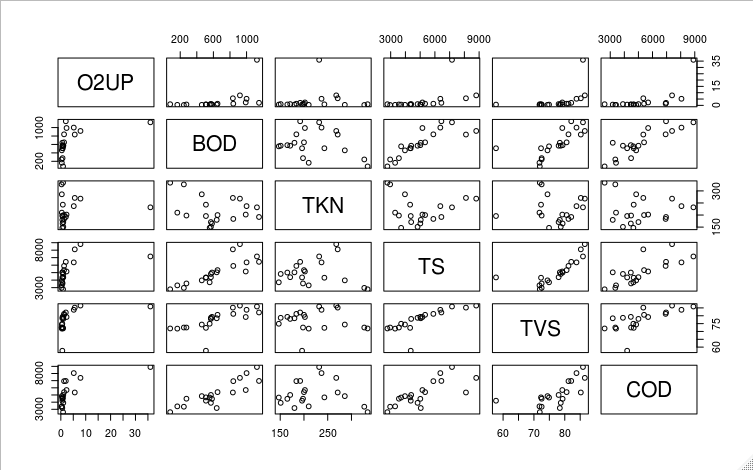
dwaste <- read.csv("/home/thomas/git/datascience/MSDataSci/MATH550/data/dwaste.csv", header=TRUE)

**# removing Day**

dwaste <- subset(dwaste, select=-c(Day))

attach(dwaste)

pairs(O2UP~., data = dwaste)



**# Already, I see that a transformation of at least the O2UP variable might be needed.**

**# This assumption is just based on visual inspection and will be investigated later.**

model <- lm(O2UP ~., data=dwaste)

summary(model)

**# Running a full summary of a linear model now does not give us the best R-squared value at 0.2609.**

**# Let's try a transform...**

Residuals:

Min 1Q Median 3Q Max

-7.6351 -2.9689 -0.4652 1.7439 20.5923

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -21.501965 23.771741 -0.905 0.3810

BOD -0.004309 0.013489 -0.319 0.7541

TKN 0.019371 0.032880 0.589 0.5651

TS 0.000189 0.002001 0.094 0.9261

TVS 0.059912 0.364318 0.164 0.8717

COD 0.003455 0.001919 1.800 0.0934 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

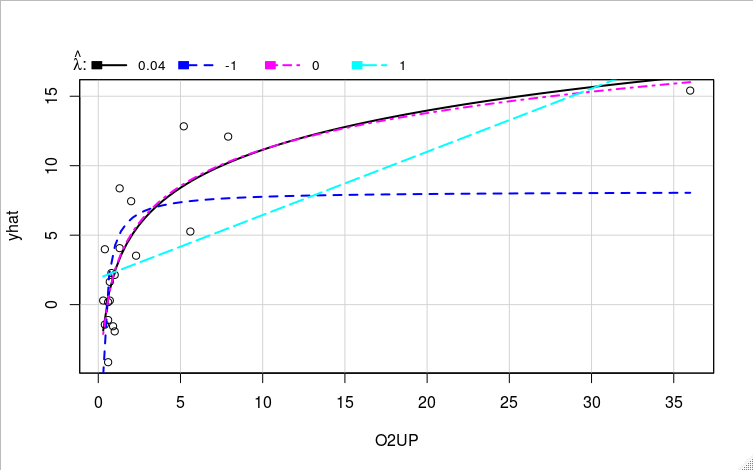
Residual standard error: 6.813 on 14 degrees of freedom

Multiple R-squared: 0.4554, Adjusted R-squared: 0.2609

F-statistic: 2.342 on 5 and 14 DF, p-value: 0.09626

par(mfrow = c(1,1))

invResults\_dwaste <- invResPlot(model)



**# The Inverse response plot picked out a visually appropriate lambda-hat of about 0.04**

**# to use as an exponent in the power transform for O2UP**

**# assigning lambda-hat to a variable**

lambdaHat\_dwaste <- invResults\_dwaste$lambda[1]

p1 <- powerTransform(O2UP ~.,data=dwaste)

**# Power transforming O2UP**

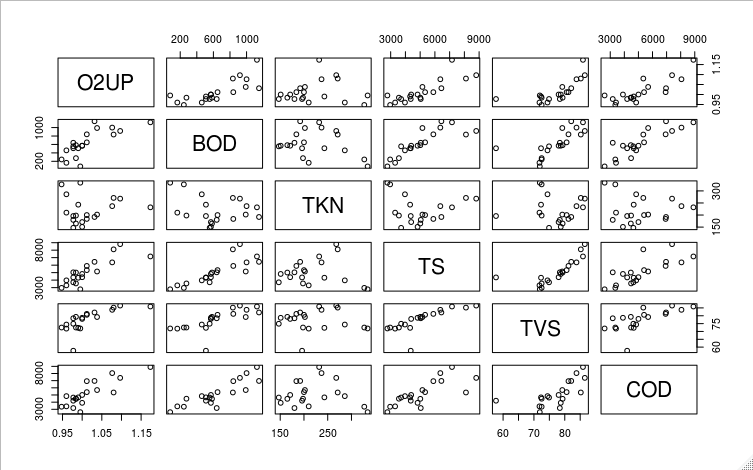
dwaste\_power <- dwaste

dwaste\_power$O2UP <- dwaste\_power$O2UP^lambdaHat\_dwaste

dwaste\_power

attach(dwaste\_power)

pairs(O2UP~., data = dwaste\_power)



**# The new pairs plot looks much better. Now, a rough linear relationship can be seen**

**# between O2UP and the predictors.**

model\_dwaste\_power <- lm(O2UP ~., data=dwaste\_power)

summary(model\_dwaste\_power)

**# The adjusted R-squared value for the new power-transformed model is a lot better at 0.7315.**

**# Log transforming O2UP**

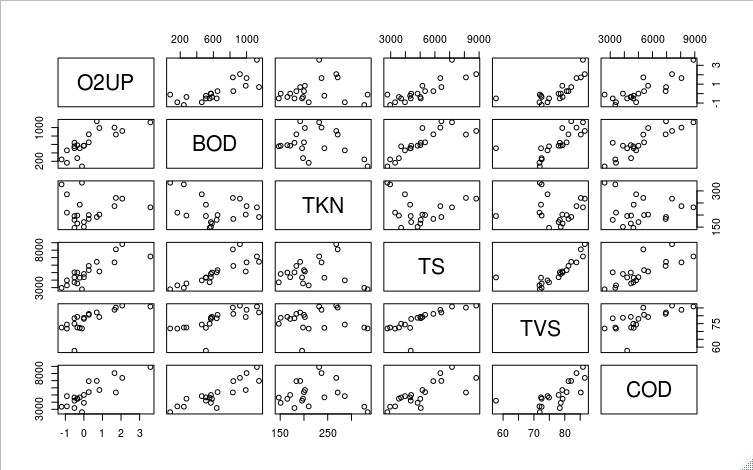
dwaste\_log <- dwaste

dwaste\_log$O2UP <- log(dwaste\_log$O2UP)

dwaste\_log

attach(dwaste\_log)

pairs(O2UP~., data = dwaste\_log)

****

**# The new pairs plot looks much better. Now, a rough linear relationship can be seen**

**# between O2UP and the predictors.**

model\_dwaste\_log <- lm(O2UP ~., data=dwaste\_log)

summary(model\_dwaste\_log)

**# The adjusted R-squared value for the new power-transformed model is a lot better at 0.7413.**

**# the log-transform performs slightly better than the power transform.**

**# I will be continuing with the log transform model.**

**# In the problem summary (and as can be seen in the new pairs plot),**

**# it alluded that there were several predictors that coud be correlated.**

**# because of this, Iwill perform backwards step AIC and BIC to reduce the**

**# number of predictors and potential multicollinearity**

n <- length(model\_dwaste\_log$residuals)

backAIC <- step(model\_dwaste\_log,direction="backward", data=dwaste\_log)

O2UP ~ TS + COD

Df Sum of Sq RSS AIC

<none> 5.7530 -18.920

- COD 1 2.3963 8.1493 -13.956

- TS 1 2.4985 8.2515 -13.707

backBIC <- step(model\_dwaste\_log,direction="backward", data=dwaste\_log, k=log(n))

O2UP ~ TS + COD

Df Sum of Sq RSS AIC

<none> 5.7530 -15.933

- COD 1 2.3963 8.1493 -11.964

- TS 1 2.4985 8.2515 -11.715

**# For both the AIC and BIC bact step methods, I am given TS and COD**

**# This simplifies the decision making for which method to choose because they both agree.**

**# AIC & BIC**

model\_dwaste\_log <- lm(O2UP~TS+COD, data=dwaste\_log)

summary(model\_dwaste\_log)

pairs(O2UP~TS+COD, data = dwaste\_log)

Residuals:

Min 1Q Median 3Q Max

-0.86663 -0.21271 -0.09741 0.14407 1.37752

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.1547134 0.4532809 -6.960 2.3e-06 \*\*\*

TS 0.0003435 0.0001264 2.717 0.0146 \*

COD 0.0003258 0.0001224 2.661 0.0165 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5817 on 17 degrees of freedom

Multiple R-squared: 0.7857, Adjusted R-squared: 0.7605

F-statistic: 31.17 on 2 and 17 DF, p-value: 2.058e-06

**# When I go back to look at the problem statement, it does not specifically mention**

**# any previously understood relationship between TS, and COD.**

**# This suggests that perhaps the span of these two predictors will still be appropriate**

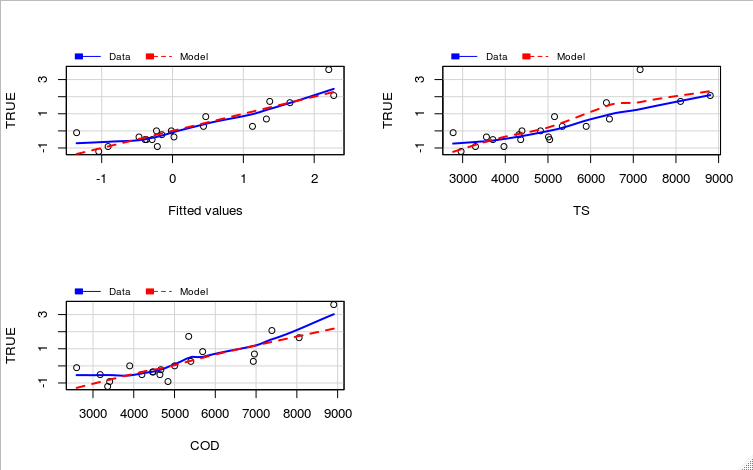
**# for a model. So, I will be using the predictors given to us by the AIC and BIC step results.**

par(mfrow = c(2,2))

mmp(model\_dwaste\_log)

mmp(model\_dwaste\_log,TS)

mmp(model\_dwaste\_log,COD)

****

**# the Marginal Model plots for the model and the individual predictors**

**# look good. The data tracks well with the proposed model**

StanRes <- rstandard(model\_dwaste\_log)

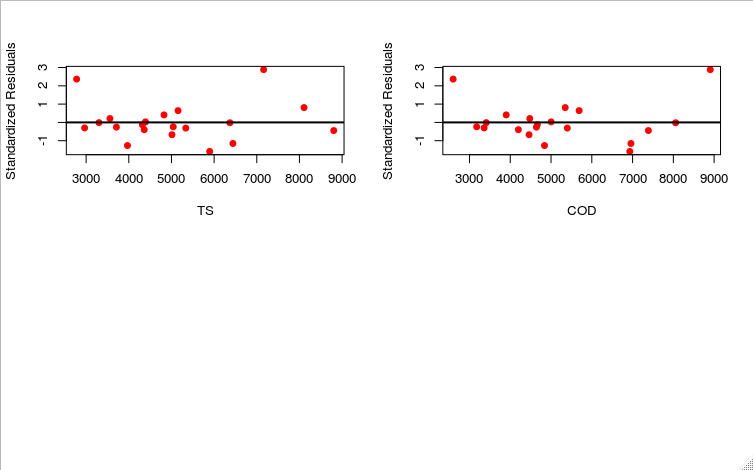
par(mfrow = c(2,2))

plot(TS,StanRes, ylab="Standardized Residuals",pch=19,col=10)

abline(h=0,lwd=2)

plot(COD,StanRes, ylab="Standardized Residuals",pch=19,col=10)

abline(h=0,lwd=2)



**# The standard residuals also are looking nice and flat.**

**# If I had to investigate the model further, I would look into Outlier Detection.**

**# There are a few points that look like they might beproviding much more leverage**

**# than others and may need to be addressed.**