Nonlinear models in R

Table of Contents

### Load libraies for plots and data analysis  
library(tidyr)  
library(ggplot2)  
library(sjPlot)

## #refugeeswelcome

library(sjmisc)

##   
## Attaching package: 'sjmisc'

## The following object is masked from 'package:tidyr':  
##   
## replace\_na

library(sjlabelled)  
  
# load response variables  
  
### Read data for specific gene expression  
dat1 <- read.csv(file="ENSG115944.csv" )  
head(dat1)

## Gene Response  
## 1 GTEX-111CU 0.39017555  
## 2 GTEX-111FC 0.03120579  
## 3 GTEX-111VG 0.19041977  
## 4 GTEX-111YS -0.43028956  
## 5 GTEX-1122O -0.38954960  
## 6 GTEX-1128S -0.61189468

# STR data for locus of interest  
dat2 <- read.csv(file="ENS115944STRup2.csv",header=FALSE )  
dat2[1:5,1:5]

## V1 V2 V3 V4 V5  
## 1 ï»¿chrom chr2 chr2 chr2 chr2  
## 2 start 42561331 42561688 42562461 42571991  
## 3 GTEX-PLZ4 0 0 <NA> 0  
## 4 GTEX-PLZ4 0 0 <NA> -1  
## 5 GTEX-NFK9 0 0 1 <NA>

# bring response and STR information together.  
dat3 <- merge(dat1,dat2[,c("V1","V42")],by.x="Gene",by.y="V1")  
dat3[1:5,]

## Gene Response V42  
## 1 GTEX-111CU 0.39017555 0  
## 2 GTEX-111CU 0.39017555 0  
## 3 GTEX-111FC 0.03120579 0  
## 4 GTEX-111FC 0.03120579 -12  
## 5 GTEX-111VG 0.19041977 13

#We only want the one locus. Drop all other levels from factor variables.  
dat4 <- droplevels(dat3)  
  
### Copy the data for detailed analsis requiring madifcations  
dat3temp <- dat3  
dat3temp$STR <- as.numeric(as.character(dat3temp$V42))  
  
dat3temp[seq(1,nrow(dat3temp),2),"side"] <- "L"  
dat3temp[seq(2,nrow(dat3temp),2),"side"] <- "R"

*In the next section we will have the following analysis*

The STR(s) are on two independent chanels. To utilize the values to determine significant effects the chanels need to be combined. The following need to be demntrated:

* The chanel has no effect (i.e. left or right), this shows the values are assigned at random.
* The value of the STR should be equal and opposite across the chanel, the values -4 and 4 cancel for example
* The STR and STR^2 are significant along with the interaction of STR and STR^2

## Model to determine if there is any significance by side , there should be none.  
sidelm <- lm(Response~side,data=dat3temp)  
summary(sidelm)

##   
## Call:  
## lm(formula = Response ~ side, data = dat3temp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.44161 -0.35593 0.00199 0.33249 1.49277   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 9.411e-03 3.239e-02 0.291 0.772  
## sideR -1.201e-16 4.581e-02 0.000 1.000  
##   
## Residual standard error: 0.5081 on 490 degrees of freedom  
## Multiple R-squared: 2.447e-32, Adjusted R-squared: -0.002041   
## F-statistic: 1.199e-29 on 1 and 490 DF, p-value: 1

## Model to determine if there is any significance by STR by itself , there should be none.  
STRlm <- lm(Response~STR,data=dat3temp)  
summary(STRlm)

##   
## Call:  
## lm(formula = Response ~ STR, data = dat3temp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.50798 -0.33355 0.00464 0.31335 1.47467   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.048965 0.026166 1.871 0.062 .  
## STR 0.005364 0.003634 1.476 0.141   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5016 on 400 degrees of freedom  
## (90 observations deleted due to missingness)  
## Multiple R-squared: 0.005417, Adjusted R-squared: 0.00293   
## F-statistic: 2.179 on 1 and 400 DF, p-value: 0.1407

## Model to determine if there is any significance by STR by itself , there should be none.  
STRsidelm <- lm(Response~STR+side,data=dat3temp)  
summary(STRsidelm)

##   
## Call:  
## lm(formula = Response ~ STR + side, data = dat3temp)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.50848 -0.33330 0.00465 0.31285 1.47516   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.0494605 0.0363177 1.362 0.174  
## STR 0.0053645 0.0036388 1.474 0.141  
## sideR -0.0009875 0.0500976 -0.020 0.984  
##   
## Residual standard error: 0.5022 on 399 degrees of freedom  
## (90 observations deleted due to missingness)  
## Multiple R-squared: 0.005418, Adjusted R-squared: 0.0004324   
## F-statistic: 1.087 on 2 and 399 DF, p-value: 0.3383

tab\_model(sidelm,STRlm,STRsidelm)

Response

Response

Response

Predictors

Estimates

CI

p

Estimates

CI

p

Estimates

CI

p

(Intercept)

0.01

-0.05 – 0.07

0.772

0.05

-0.00 – 0.10

0.062

0.05

-0.02 – 0.12

0.174

side [R]

-0.00

-0.09 – 0.09

1.000

-0.00

-0.10 – 0.10

0.984

STR

0.01

-0.00 – 0.01

0.141

0.01

-0.00 – 0.01

0.141

Observations

492

402

402

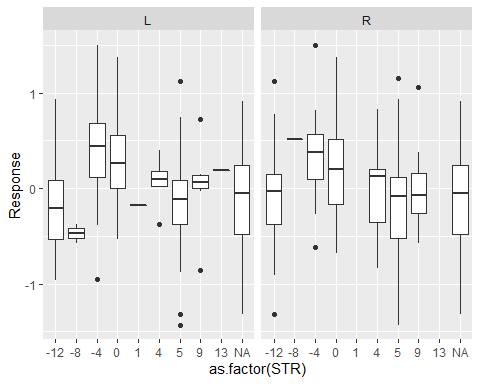
R2 / R2 adjusted

0.000 / -0.002

0.005 / 0.003

0.005 / 0.000

p1 <- ggplot(dat3temp, aes(as.factor(STR),Response)) + facet\_grid(~side) + geom\_boxplot()  
  
  
p1



The STR(s) chanels are combined “summed” across.

* The chanel has no effect (i.e. left or right), this shows the values are assigned at random.
* The value of the STR should be equal and opposite across the chanel, the values -4 and 4 cancel for example
* The STR and STR^2 are significant along with the interaction of STR and STR^2

### V42 is the start postion we want. start = 42625365  
  
#### Want the STR value to numeric not factor  
dat4$STR <- as.numeric(as.character(dat4$V42))  
dat4$STRquad <- dat4$STR\*\*2  
  
### Add L/R side can be aribtrary becuase of the model steps above.  
### i.e no signifcance when choosing side or STR just need to constent  
dat4[seq(1,nrow(dat4),2),"side"] <- "L"  
dat4[seq(2,nrow(dat4),2),"side"] <- "R"  
  
dat4[seq(1,nrow(dat4),2),"opside"] <- "R"  
dat4[seq(2,nrow(dat4),2),"opside"] <- "L"  
  
### Model with sqaured term.  
seplm4a <- lm(dat4$Response ~ dat4$STR + dat4$STRquad + dat4$side )  
  
## STR matter and the squre matters , side L/R does not.  
summary(seplm4a)

##   
## Call:  
## lm(formula = dat4$Response ~ dat4$STR + dat4$STRquad + dat4$side)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.43427 -0.33825 0.01143 0.29074 1.37093   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.206429 0.041478 4.977 9.64e-07 \*\*\*  
## dat4$STR -0.020483 0.005139 -3.986 8.00e-05 \*\*\*  
## dat4$STRquad -0.004078 0.000601 -6.785 4.22e-11 \*\*\*  
## dat4$sideR -0.001241 0.047489 -0.026 0.979   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.476 on 398 degrees of freedom  
## (90 observations deleted due to missingness)  
## Multiple R-squared: 0.1085, Adjusted R-squared: 0.1018   
## F-statistic: 16.15 on 3 and 398 DF, p-value: 6.316e-10

BIC(seplm4a)

## [1] 570.009

seplm4b <- lm(dat4$Response ~ dat4$STR)  
summary(seplm4b)

##   
## Call:  
## lm(formula = dat4$Response ~ dat4$STR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.50798 -0.33355 0.00464 0.31335 1.47467   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.048965 0.026166 1.871 0.062 .  
## dat4$STR 0.005364 0.003634 1.476 0.141   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5016 on 400 degrees of freedom  
## (90 observations deleted due to missingness)  
## Multiple R-squared: 0.005417, Adjusted R-squared: 0.00293   
## F-statistic: 2.179 on 1 and 400 DF, p-value: 0.1407

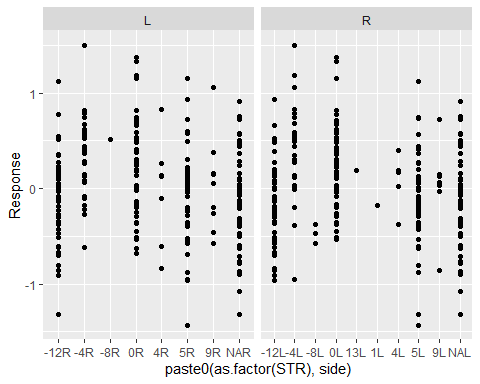
BIC(seplm4b)

## [1] 602.0154

summary(lm(dat4$Response ~ dat4$STR + abs(dat4$STR)))

##   
## Call:  
## lm(formula = dat4$Response ~ dat4$STR + abs(dat4$STR))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.40242 -0.33603 0.00716 0.29906 1.36712   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.264218 0.040033 6.600 1.31e-10 \*\*\*  
## dat4$STR -0.013255 0.004386 -3.022 0.00268 \*\*   
## abs(dat4$STR) -0.045545 0.006652 -6.847 2.87e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4751 on 399 degrees of freedom  
## (90 observations deleted due to missingness)  
## Multiple R-squared: 0.11, Adjusted R-squared: 0.1055   
## F-statistic: 24.65 on 2 and 399 DF, p-value: 8.041e-11

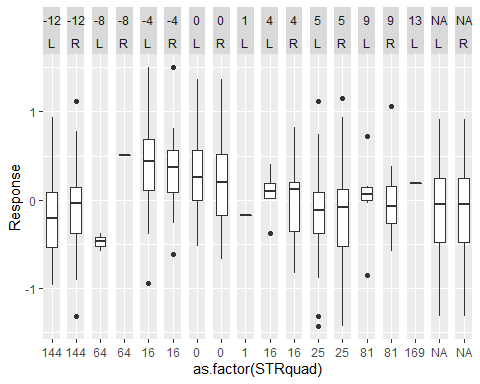
#dat4$V42 <- factor(dat4$V42,levels(dat4$V42)[order(abs(dat4$STR))])  
p2a <- ggplot(dat4, aes(paste0(as.factor(STR),side),Response)) + geom\_point() + facet\_grid(~opside,scales="free\_x")  
  
p2a



summary(lm(Response~paste0(as.factor(STRquad),side),data = dat4))

##   
## Call:  
## lm(formula = Response ~ paste0(as.factor(STRquad), side), data = dat4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.29810 -0.32391 0.01027 0.27038 1.28309   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.27822 0.06414 4.338 1.76e-05 \*\*\*  
## paste0(as.factor(STRquad), side)0R -0.07624 0.09201 -0.829 0.40772   
## paste0(as.factor(STRquad), side)144L -0.46205 0.09071 -5.094 5.07e-07 \*\*\*  
## paste0(as.factor(STRquad), side)144R -0.40253 0.09156 -4.396 1.36e-05 \*\*\*  
## paste0(as.factor(STRquad), side)169L -0.08780 0.47998 -0.183 0.85494   
## paste0(as.factor(STRquad), side)16L 0.04910 0.10683 0.460 0.64601   
## paste0(as.factor(STRquad), side)16R -0.02035 0.10114 -0.201 0.84061   
## paste0(as.factor(STRquad), side)1L -0.45599 0.47998 -0.950 0.34259   
## paste0(as.factor(STRquad), side)25L -0.41517 0.09396 -4.419 1.23e-05 \*\*\*  
## paste0(as.factor(STRquad), side)25R -0.41232 0.09396 -4.388 1.41e-05 \*\*\*  
## paste0(as.factor(STRquad), side)64L -0.75203 0.28202 -2.667 0.00792 \*\*   
## paste0(as.factor(STRquad), side)64R 0.23091 0.47998 0.481 0.63068   
## paste0(as.factor(STRquad), side)81L -0.24819 0.19089 -1.300 0.19417   
## paste0(as.factor(STRquad), side)81R -0.27440 0.16353 -1.678 0.09400 .   
## paste0(as.factor(STRquad), side)NAL -0.39488 0.09561 -4.130 4.29e-05 \*\*\*  
## paste0(as.factor(STRquad), side)NAR -0.39488 0.09561 -4.130 4.29e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4757 on 476 degrees of freedom  
## Multiple R-squared: 0.1485, Adjusted R-squared: 0.1217   
## F-statistic: 5.536 on 15 and 476 DF, p-value: 1.936e-10

p2 <- ggplot(dat4, aes(as.factor(STRquad),Response)) + geom\_boxplot() + facet\_grid(~STR+side,scales="free\_x")  
  
p2



dat5 <- aggregate(dat4[,c("STR","STRquad")],by=list(dat4$Gene,dat4$Response),FUN=sum,na.rm=TRUE)  
  
names(dat5) <- c("Gene","y","STR","STRsqr")  
  
  
agglm <- lm(y ~ STR+STR\*STRsqr,dat=dat5)  
summary(agglm)

##   
## Call:  
## lm(formula = y ~ STR + STR \* STRsqr, data = dat5)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.46469 -0.32916 0.03344 0.30942 1.22551   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.458e-01 4.533e-02 3.216 0.001475 \*\*   
## STR -2.701e-02 7.974e-03 -3.387 0.000826 \*\*\*  
## STRsqr -1.724e-03 5.999e-04 -2.874 0.004416 \*\*   
## STR:STRsqr 8.863e-05 4.182e-05 2.119 0.035090 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4861 on 242 degrees of freedom  
## Multiple R-squared: 0.09588, Adjusted R-squared: 0.08467   
## F-statistic: 8.555 on 3 and 242 DF, p-value: 2.02e-05

SSTotal <- var( dat5$y ) \* (nrow(dat5)-1)  
 SSE <- sum( agglm$resid^2 )  
SSreg <- SSTotal - SSE  
  
SSTotal

## [1] 63.24711

SSreg

## [1] 6.064179

SSE

## [1] 57.18293

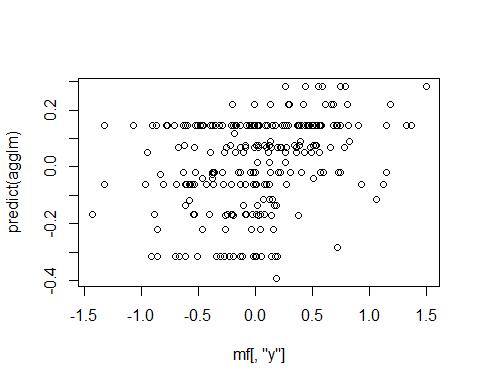
BIC(agglm)

## [1] 366.7117

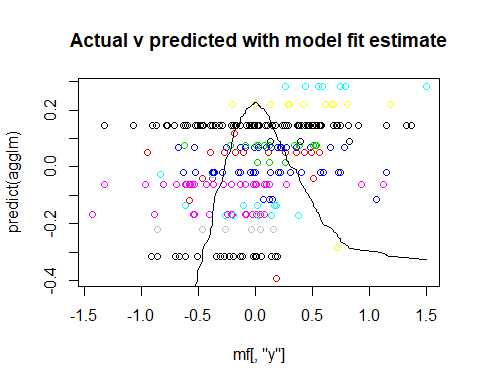
sum(resid(agglm)^2)

## [1] 57.18293

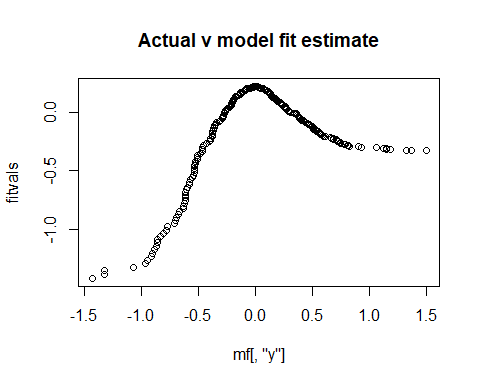
mf <- model.frame(agglm)  
plot(mf[,"y"],predict(agglm))



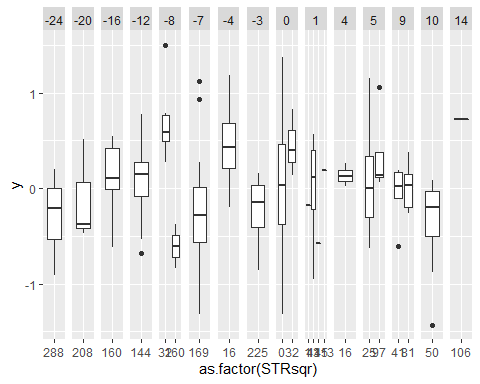
xvals <- seq(-24,14,length.out=246)  
fitvals <- 0.1458 -0.02701\*xvals - 0.001724\*xvals\*\*2 + 0.00008863\*(xvals\*xvals\*\*2)  
  
  
plot(mf[,"y"],predict(agglm),col=as.factor(mf[,"STR"]),main="Actual v predicted with model fit estimate")  
lines(mf[,"y"],fitvals)



plot(mf[,"y"],fitvals,main="Actual v model fit estimate")



#Coefficients:  
# Estimate Std. Error t value Pr(>|t|)   
#(Intercept) 1.458e-01 4.533e-02 3.216 0.001475 \*\*   
#STR -2.701e-02 7.974e-03 -3.387 0.000826 \*\*\*  
#STRsqr -1.724e-03 5.999e-04 -2.874 0.004416 \*\*   
#STR:STRsqr 8.863e-05 4.182e-05 2.119 0.035090 \*   
  
  
p3 <- ggplot(dat5, aes(as.factor(STRsqr),y)) + geom\_boxplot() + facet\_grid(~STR,scales="free\_x")  
  
p3



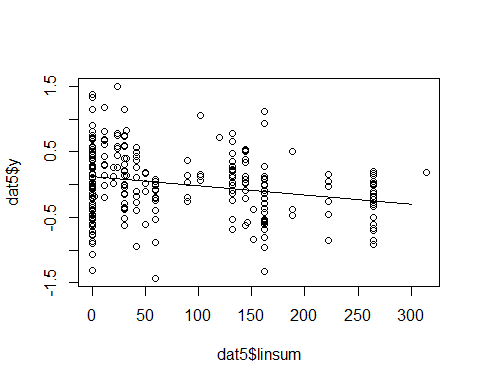
#agglmcat <- lm(dat5$Group.2~ as.factor(dat5$STR)+as.factor(dat5$STRquad))  
#summary(agglmcat)  
  
summary(lm(dat5$y ~ dat5$STR))

##   
## Call:  
## lm(formula = dat5$y ~ dat5$STR)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.49266 -0.33472 0.00508 0.31455 1.51003   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 0.022510 0.034610 0.650 0.516  
## dat5$STR 0.003795 0.003538 1.073 0.284  
##   
## Residual standard error: 0.5079 on 244 degrees of freedom  
## Multiple R-squared: 0.004693, Adjusted R-squared: 0.0006143   
## F-statistic: 1.151 on 1 and 244 DF, p-value: 0.2845

dat5$linsum <- dat5$STR + dat5$STRsqr  
dat5$linsumint <- dat5$STR + dat5$STRsqr + (dat5$STR\*dat5$STRsqr)  
  
linmodsqr <- lm(dat5$y ~ dat5$linsum)  
  
BIC(linmodsqr)

## [1] 366.159

#(Intercept) 0.1240251 0.0434871 2.852 0.004717 \*\*   
#dat5$linsum -0.0014102 0.0003685 -3.827 0.000165 \*\*\*  
  
xvals <- seq(0,300,1)  
  
plot(dat5$linsum,dat5$y)  
lines(xvals,0.124-0.00141\*xvals)



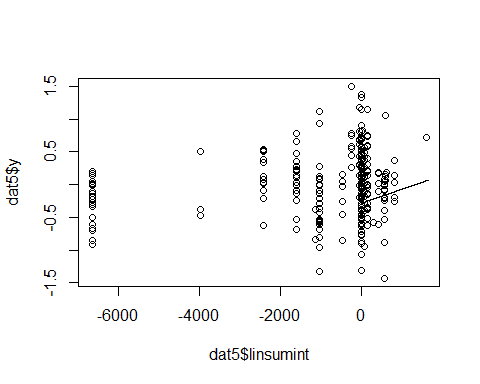
linmodsqrint <- lm(dat5$y ~ dat5$linsumint)  
  
BIC(linmodsqrint)

## [1] 373.9619

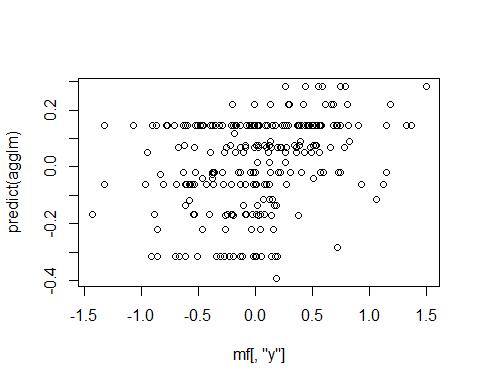
summary(linmodsqrint)

##   
## Call:  
## lm(formula = dat5$y ~ dat5$linsumint)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.50233 -0.34142 0.00054 0.33172 1.46588   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.621e-02 3.510e-02 1.316 0.189   
## dat5$linsumint 4.273e-05 1.667e-05 2.563 0.011 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.5024 on 244 degrees of freedom  
## Multiple R-squared: 0.02621, Adjusted R-squared: 0.02222   
## F-statistic: 6.568 on 1 and 244 DF, p-value: 0.01098

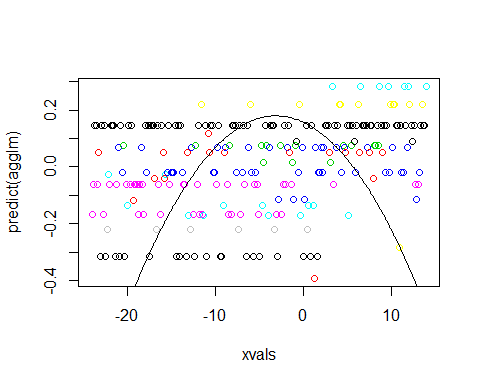
xvals <- seq(-6650,1600,5)  
  
plot(dat5$linsumint,dat5$y)  
lines(x=0.0000427\*xvals)



mf <- model.frame(agglm)  
plot(mf[,"y"],predict(agglm))



xvals <- seq(-24,14,length.out=246)  
fitvals <- 0.158 -0.0144\*xvals - 0.00234\*xvals\*\*2  
plot(xvals,predict(agglm),col=as.factor(mf[,"STR"]))  
lines(xvals,fitvals)



Add a new chunk by clicking the *Insert Chunk* button on the toolbar or by pressing *Ctrl+Alt+I*.

When you save the notebook, an HTML file containing the code and output will be saved alongside it (click the *Preview* button or press *Ctrl+Shift+K* to preview the HTML file).

The preview shows you a rendered HTML copy of the contents of the editor. Consequently, unlike *Knit*, *Preview* does not run any R code chunks. Instead, the output of the chunk when it was last run in the editor is displayed.