Parent and Provider Perceptions of Behavioral Healthcare in Pediatric Primary Care (PI: Andrew Riley; BDP2-262)

Benjamin Chan (chanb@ohsu.edu)

2018-06-07

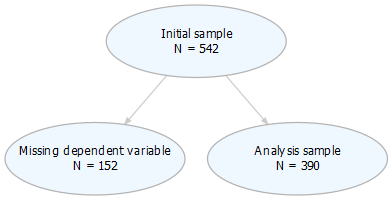
# Import Andrew’s SPSS data

Map new names to variables.

|  |  |
| --- | --- |
| oldnames | newnames |
| record\_id | id |
| eng\_span | languageSurvey |
| children\_totv\_1 | totalChildren |
| oldest\_middle\_youngest | birthOrder |
| child\_age\_years | childAge |
| Age\_Dichot | childAgeDichotomous |
| child\_sexv\_1 | childSex |
| child\_ethnicity | childEthnicity |
| child\_racev\_1\_\_\_1 | childRaceWhite |
| child\_racev\_1\_\_\_2 | childRaceAsian |
| child\_racev\_1\_\_\_3 | childRaceAfrAm |
| child\_racev\_1\_\_\_4 | childRaceAIAN |
| child\_racev\_1\_\_\_5 | childRaceNHPI |
| child\_racev\_1\_\_\_6 | childRaceOther |
| child\_racev\_1\_\_\_7 | childRaceNoResp |
| visit\_typev\_1 | visitType |
| related\_child | childRelationship |
| gender | parentGender |
| parent\_sexv\_1 | parentSex |
| parent\_agev\_1 | parentAge |
| parent\_ethnicity | parentEthnicity |
| parent\_race\_\_\_1 | parentRaceWhite |
| parent\_race\_\_\_2 | parentRaceAsian |
| parent\_race\_\_\_3 | parentRaceAfrAm |
| parent\_race\_\_\_4 | parentRaceAIAN |
| parent\_race\_\_\_5 | parentRaceNHPI |
| parent\_race\_\_\_6 | parentRaceOther |
| parent\_race\_\_\_7 | parentRaceNoResp |
| marital\_status | parentMaritalStatus |
| parenting\_situationv\_1 | parentSituation |
| zipcode | zipcode |
| community\_type | community |
| distance | distance |
| parent\_educationv\_1 | parentEducation |
| annual\_income | income |
| internet | internet |
| ECBI\_intensity\_raw\_score | ECBI\_intensity\_raw\_score |
| ECBI\_intensity\_T\_score | ECBI\_intensity\_T\_score |
| ECBI\_intensity\_clinical\_cutoff | ECBI\_intensity\_clinical\_cutoff |
| ECBI\_problem\_raw\_score | ECBI\_problem\_raw\_score |
| ECBI\_problem\_T\_score | ECBI\_problem\_T\_score |
| ECBI\_problem\_clinical\_cutoff | ECBI\_problem\_clinical\_cutoff |
| ECBI\_OPP\_Tot | ECBI\_OPP\_Tot |
| ECBI\_Inatt\_Tot | ECBI\_Inatt\_Tot |
| ECBI\_Cond\_Tot | ECBI\_Cond\_Tot |
| MAPS\_PP | MAPS\_PP |
| MAPS\_PR | MAPS\_PR |
| MAPS\_WM | MAPS\_WM |
| MAPS\_SP | MAPS\_SP |
| MAPS\_HS | MAPS\_HS |
| MAPS\_LC | MAPS\_LC |
| MAPS\_PC | MAPS\_PC |
| MAPS\_POS | MAPS\_POS |
| MAPS\_NEG | MAPS\_NEG |
| SEPTI\_nurturance | SEPTI\_nurturance |
| SEPTI\_n\_clinical\_cutoff | SEPTI\_n\_clinical\_cutoff |
| SEPTI\_discipline | SEPTI\_discipline |
| SEPTI\_d\_clinical\_cutoff | SEPTI\_d\_clinical\_cutoff |
| SEPTI\_play | SEPTI\_play |
| SEPTI\_p\_clinical\_cutoff | SEPTI\_p\_clinical\_cutoff |
| SEPTI\_routine | SEPTI\_routine |
| SEPTI\_r\_clinical\_cutoff | SEPTI\_r\_clinical\_cutoff |
| SEPTI\_total | SEPTI\_total |
| SEPTI\_total\_clin\_cutoff | SEPTI\_total\_clin\_cutoff |
| Y1 | Y1 |
| Y2 | Y2 |
| Y3 | Y3 |

Output to [CSV file](../data/processed/dataframe.csv).

Build analysis data set. Exclude if missing any dependent variable, Y1, Y2, Y3.



figures/flowChart.png

# Preprocess data

Initial preprocesssing that needs to be done that is common to Y1, Y2, and Y3.

Split data set into 70:30 training:validation samples.

inTrain <- createDataPartition(df$id, p = 0.7)  
dfTrain <- df[inTrain$Resample1, ]  
dfValid <- df[-inTrain$Resample1, ]

Preprocess the training sample.

1. Exclude near-zero variance predictors
2. Impute missing values using k-nearest neighbor

message(sprintf("Number of complete cases before imputation = %d",  
 complete.cases(dfTrain) %>% sum()))

## Number of complete cases before imputation = 247

nzv <-   
 dfTrain %>%   
 select(-c(id, Y1, Y2, Y3)) %>%   
 nearZeroVar(names = TRUE, saveMetric = TRUE) %>%  
 mutate(varname = row.names(.)) %>%   
 filter(nzv == TRUE) %>%   
 select(varname, freqRatio, percentUnique, zeroVar, nzv)   
nzv %>% kable()

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| varname | freqRatio | percentUnique | zeroVar | nzv |
| languageSurvey | 53.80000 | 0.729927 | FALSE | TRUE |
| childRaceAfrAm | 26.40000 | 0.729927 | FALSE | TRUE |
| childRaceAIAN | 38.14286 | 0.729927 | FALSE | TRUE |
| childRaceNHPI | 44.66667 | 0.729927 | FALSE | TRUE |
| childRaceOther | 21.83333 | 0.729927 | FALSE | TRUE |
| parentRaceAfrAm | 44.66667 | 0.729927 | FALSE | TRUE |
| parentRaceAIAN | 29.44444 | 0.729927 | FALSE | TRUE |
| parentRaceNHPI | 44.66667 | 0.729927 | FALSE | TRUE |
| parentRaceOther | 20.07692 | 0.729927 | FALSE | TRUE |
| internet | 33.25000 | 0.729927 | FALSE | TRUE |

dfTrainPreProc1 <-  
 dfTrain %>%   
 select(-one\_of(nzv$varname))  
dfOutcomes <-   
 dfTrainPreProc1 %>%   
 select(c(id, Y1, Y2, Y3))  
dfTrainPreProc2 <-   
 dfTrainPreProc1 %>%   
 select(-c(id, Y1, Y2, Y3))  
preProc <-  
 dfTrainPreProc2 %>%   
 preProcess(method = c("nzv", "corr", "knnImpute"), verbose = TRUE)

## 2 highly correlated predictors were removed.  
## Calculating 30 means for centering  
## Calculating 30 standard deviations for scaling

preProc

## Created from 247 samples and 53 variables  
##   
## Pre-processing:  
## - centered (30)  
## - ignored (21)  
## - 5 nearest neighbor imputation (30)  
## - removed (2)  
## - scaled (30)

dfTrainPreProc3 <-   
 predict(preProc, dfTrainPreProc2) %>%   
 mutate(childAgeDichotomous = case\_when(is.na(childAgeDichotomous) & childAge < 3 ~ 1,  
 is.na(childAgeDichotomous) & childAge >= 3 ~ 2,  
 TRUE ~ as.numeric(childAgeDichotomous))) %>%   
 mutate(childAgeDichotomous = factor(childAgeDichotomous,   
 levels = seq(2),   
 labels = c("Under 3", "3 or older")))  
dfTrainPreProc <- bind\_cols(dfOutcomes, dfTrainPreProc3)  
message(sprintf("Number of complete cases after imputation = %d",  
 complete.cases(dfTrainPreProc) %>% sum()))

## Number of complete cases after imputation = 274

dfTrainPreProc %>% write.csv("data/processed/dfTrainPreProc.csv", row.names = FALSE)  
rm(dfTrainPreProc1, dfTrainPreProc2, dfTrainPreProc3)

Set the control parameters.

ctrl <- trainControl(method = "repeatedcv",  
 number = 10,  
 repeats = 25,  
 savePredictions = TRUE,  
 allowParallel = TRUE,  
 search = "random")

Set the model and tuning parameter grid.

library(randomForest)

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

library(randomForestExplainer)  
method <- "rf"  
grid <- expand.grid(mtry = seq(5, 10, 1))

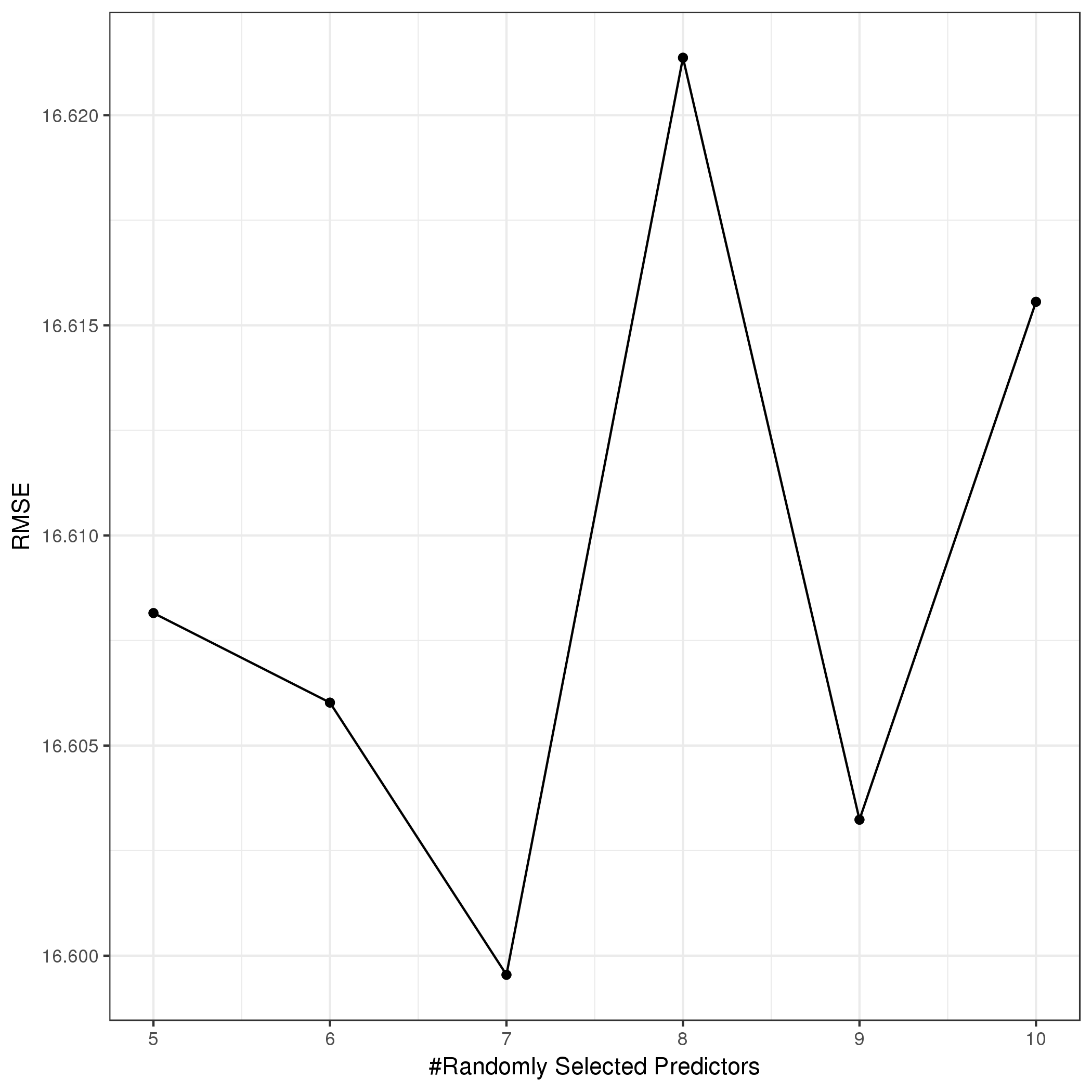
# Model 1

Prediction model for Y1.

Train model over the tuning parameters.

## Random Forest   
##   
## 274 samples  
## 51 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 25 times)   
## Summary of sample sizes: 247, 246, 247, 247, 247, 246, ...   
## Resampling results across tuning parameters:  
##   
## mtry RMSE Rsquared MAE   
## 5 16.60815 0.1469653 13.44393  
## 6 16.60602 0.1437086 13.43946  
## 7 16.59955 0.1426356 13.43343  
## 8 16.62137 0.1394873 13.44105  
## 9 16.60324 0.1415267 13.42774  
## 10 16.61556 0.1379677 13.43435  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was mtry = 7.

## Saving 7 x 7 in image  
## Saving 7 x 7 in image



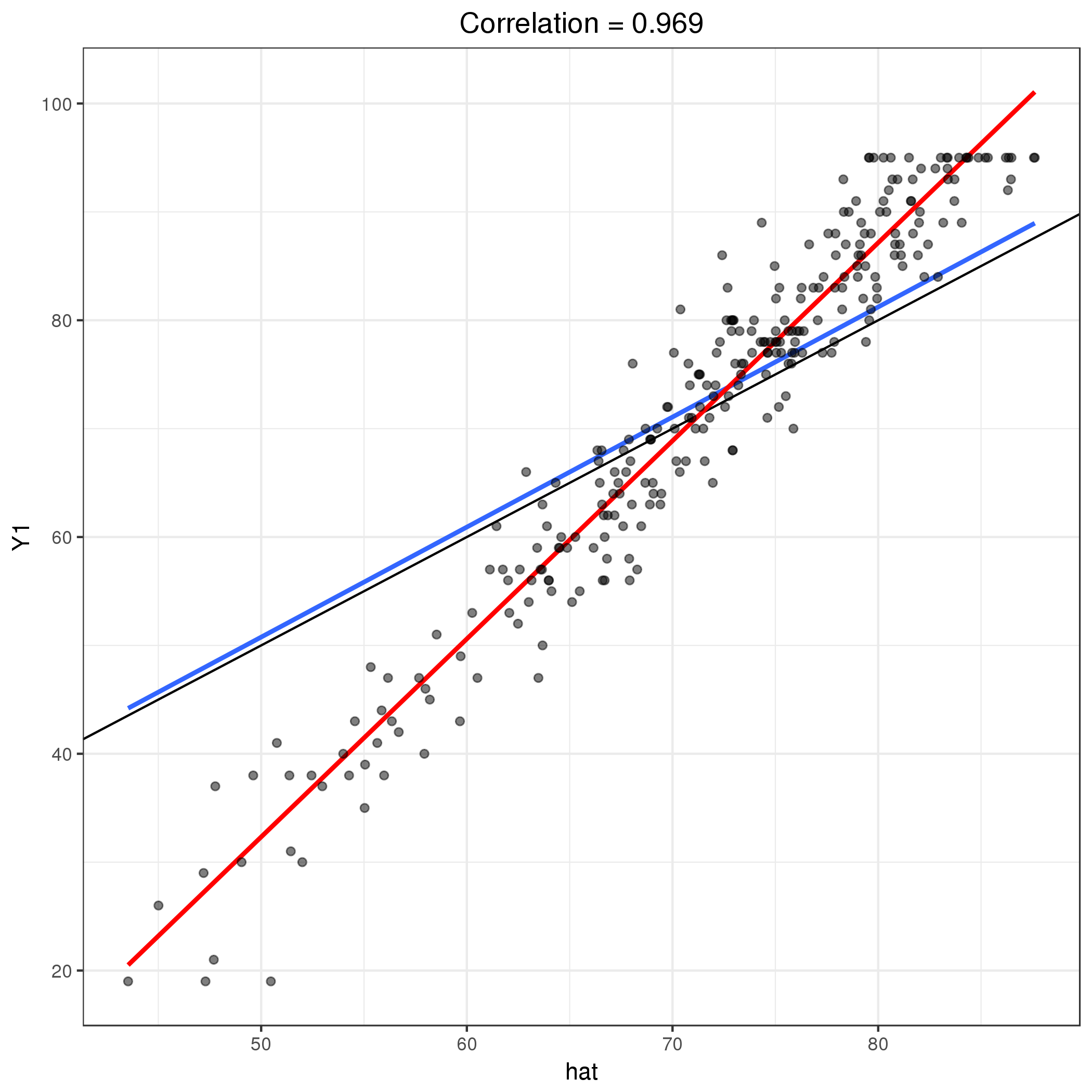
figures/Y1Training.png

##   
## Call:  
## randomForest(x = x, y = y, mtry = param$mtry, localImp = TRUE, nthreads = 8)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 7  
##   
## Mean of squared residuals: 281.8461  
## % Var explained: 9.02

## rf variable importance  
##   
## only 20 most important variables shown (out of 163)  
##   
## Overall  
## childRaceWhite 100.00  
## MAPS\_POS 88.73  
## SEPTI\_r\_clinical\_cutoff 84.26  
## parentRaceWhite 83.22  
## SEPTI\_total 79.26  
## MAPS\_SP 76.20  
## zipcode97225 72.97  
## MAPS\_PP 72.68  
## childAgeDichotomous3 or older 70.98  
## ECBI\_problem\_raw\_score 68.82  
## MAPS\_NEG 68.29  
## parentEthnicityNot Hispanic/Latino 68.19  
## parentSituationCo-parenting in separate households 67.23  
## MAPS\_LC 65.48  
## zipcode97702 63.66  
## childEthnicityNot Hispanic/Latino 62.83  
## MAPS\_HS 62.53  
## SEPTI\_discipline 62.50  
## SEPTI\_n\_clinical\_cutoff 60.93  
## SEPTI\_total\_clin\_cutoff 60.25

## RMSE Rsquared MAE   
## 8.8726672 0.9383094 7.0898860

## Y1 hat  
## Y1 1.0000000 0.9686637  
## hat 0.9686637 1.0000000

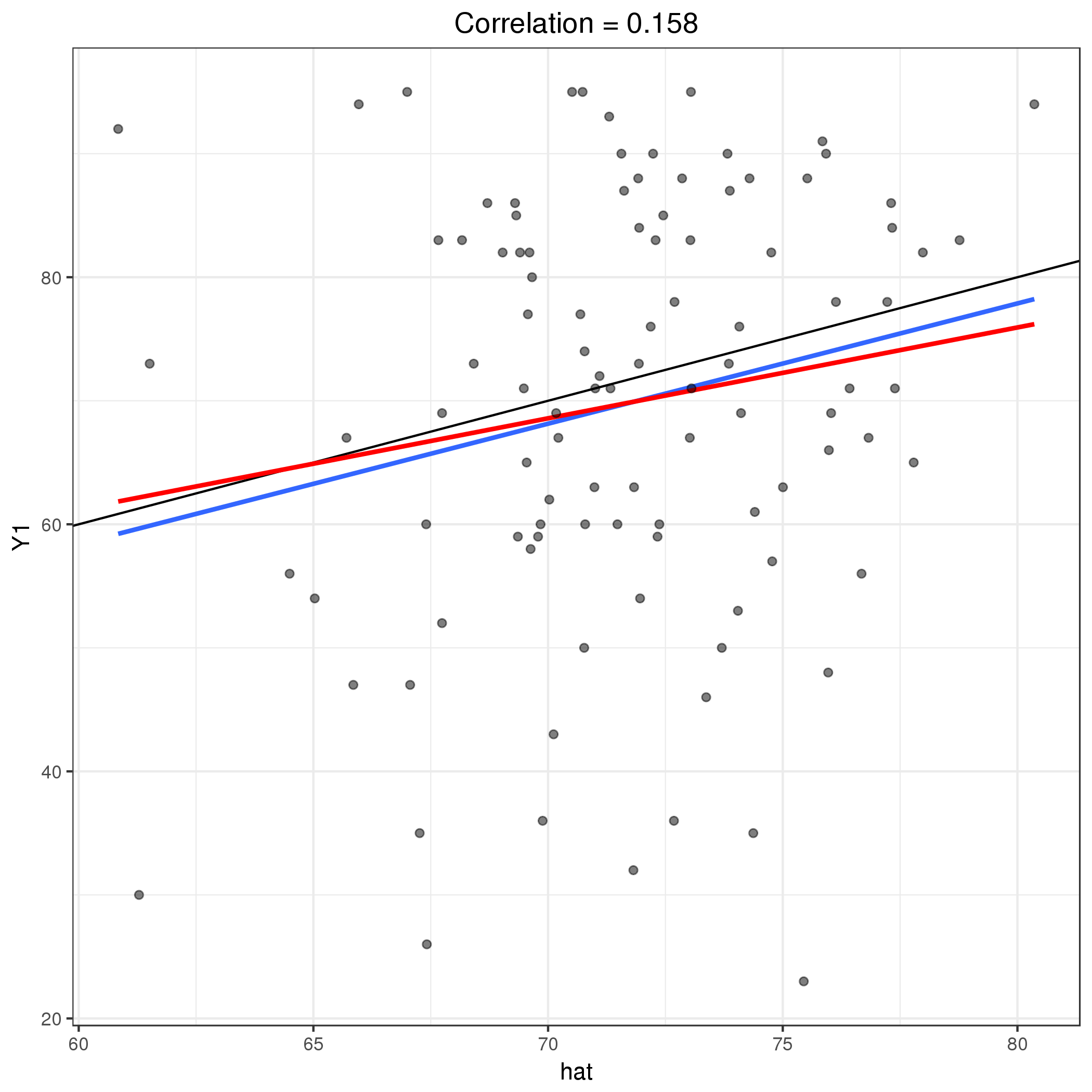


plot of chunk Y1Training-predict

Evaluate model on the validation sample.

## RMSE Rsquared MAE   
## 17.24798835 0.02505421 13.84977108

## Y1 hat  
## Y1 1.0000000 0.1582852  
## hat 0.1582852 1.0000000



plot of chunk Y1Validation-predict

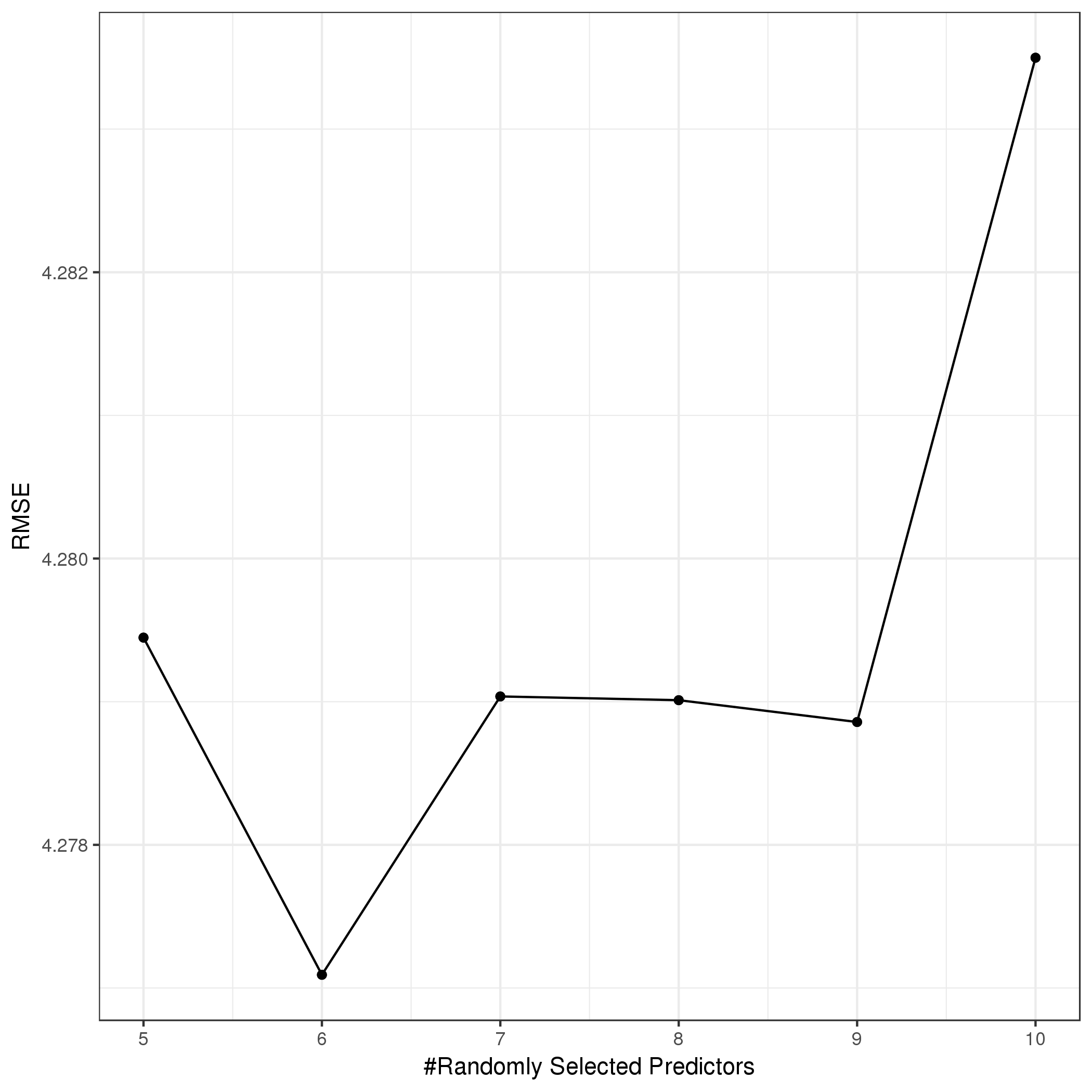
# Model 2

Prediction model for Y2.

Train model over the tuning parameters.

## Random Forest   
##   
## 274 samples  
## 51 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 25 times)   
## Summary of sample sizes: 247, 246, 245, 247, 247, 248, ...   
## Resampling results across tuning parameters:  
##   
## mtry RMSE Rsquared MAE   
## 5 4.279448 0.08849013 3.281976  
## 6 4.277092 0.08864875 3.280306  
## 7 4.279037 0.08667457 3.281294  
## 8 4.279010 0.08679174 3.280949  
## 9 4.278858 0.08764245 3.280722  
## 10 4.283499 0.08579369 3.285062  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was mtry = 6.

## Saving 7 x 7 in image  
## Saving 7 x 7 in image



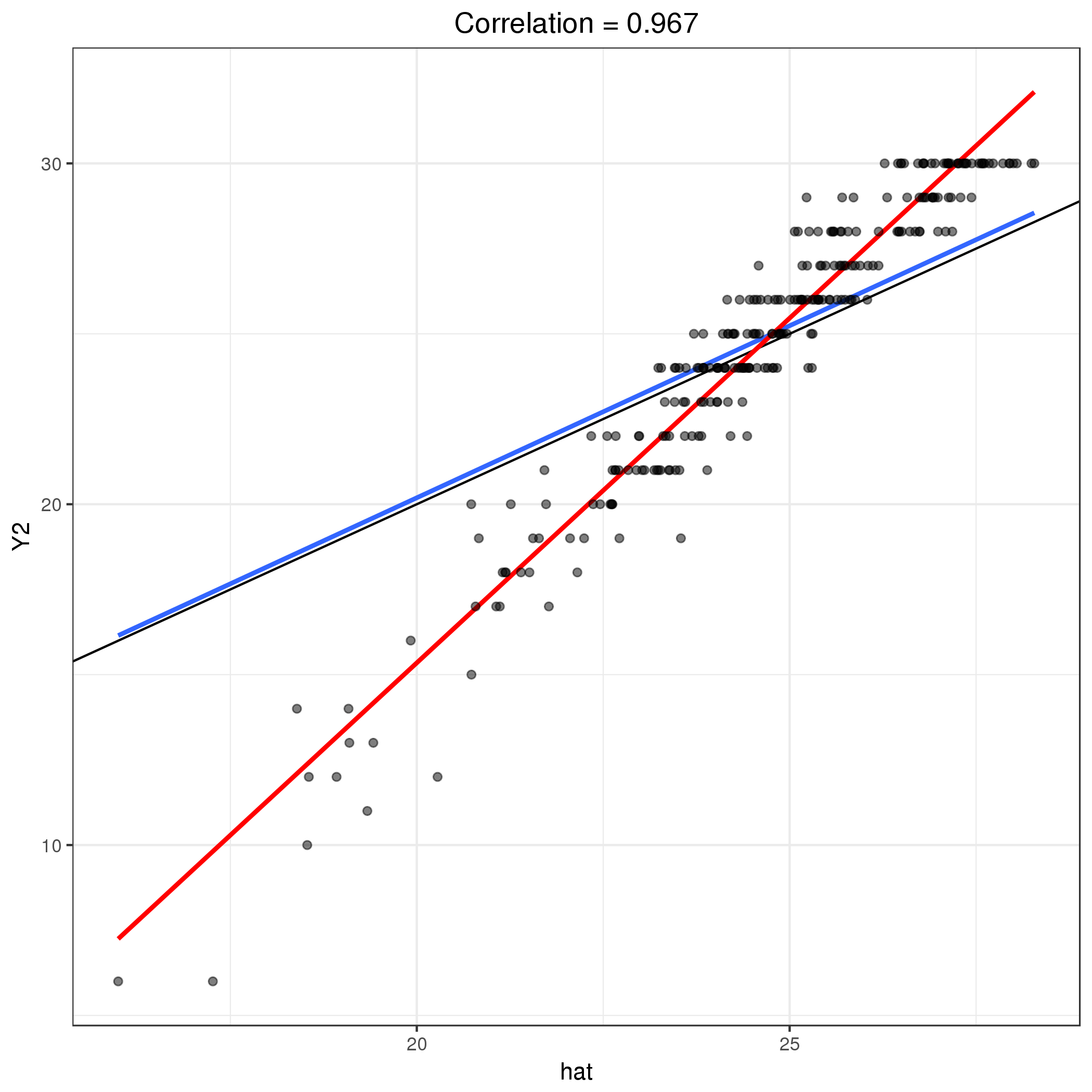
figures/Y2Training.png

##   
## Call:  
## randomForest(x = x, y = y, mtry = param$mtry, localImp = TRUE, nthreads = 8)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 6  
##   
## Mean of squared residuals: 19.02295  
## % Var explained: 3.99

## rf variable importance  
##   
## only 20 most important variables shown (out of 163)  
##   
## Overall  
## SEPTI\_r\_clinical\_cutoff 100.00  
## MAPS\_POS 90.81  
## parentSexMale 85.25  
## MAPS\_SP 84.88  
## MAPS\_PC 79.50  
## zipcode97702 78.74  
## SEPTI\_total 78.09  
## zipcode97760 77.71  
## parentSituationCo-parenting in separate households 77.49  
## SEPTI\_routine 75.27  
## parentEducationGraduate/professional school 74.65  
## zipcode97211 74.07  
## childEthnicityUnknown 71.94  
## childAge 70.15  
## parentRaceWhite 69.19  
## MAPS\_PP 69.02  
## ECBI\_problem\_clinical\_cutoff 67.52  
## income$25,001-$49,999 66.98  
## zipcode97008 66.79  
## SEPTI\_n\_clinical\_cutoff 64.47

## RMSE Rsquared MAE   
## 2.4569309 0.9343127 1.8447587

## Y2 hat  
## Y2 1.0000000 0.9665985  
## hat 0.9665985 1.0000000

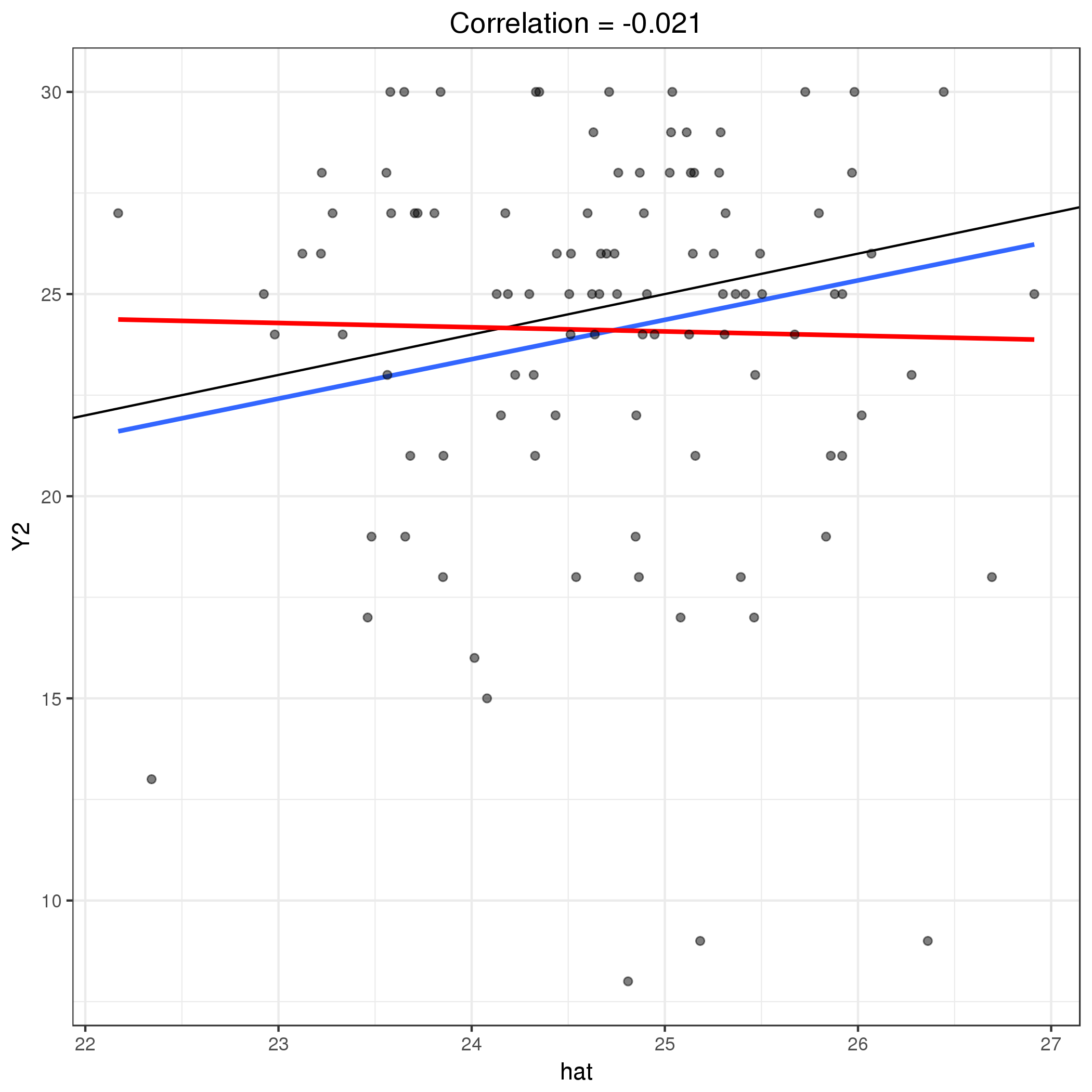


plot of chunk Y2Training-predict

Evaluate model on the validation sample.

## RMSE Rsquared MAE   
## 4.8084129417 0.0004435993 3.5358737108

## Y2 hat  
## Y2 1.0000000 -0.0210618  
## hat -0.0210618 1.0000000



plot of chunk Y2Validation-predict

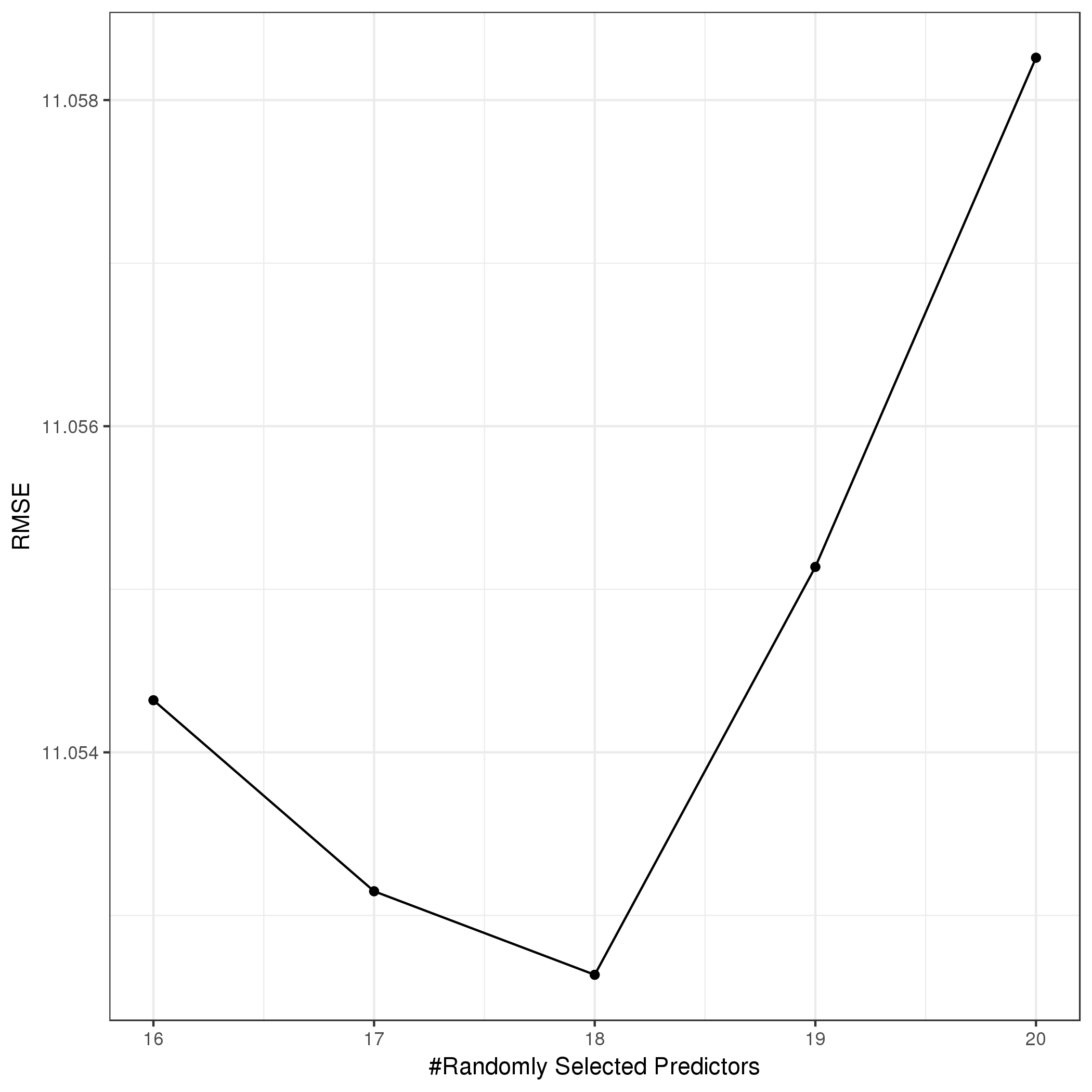
# Model 3

Prediction model for Y3.

Train model over the tuning parameters.

## Random Forest   
##   
## 274 samples  
## 51 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 25 times)   
## Summary of sample sizes: 246, 247, 245, 247, 247, 247, ...   
## Resampling results across tuning parameters:  
##   
## mtry RMSE Rsquared MAE   
## 16 11.05432 0.1509766 9.072579  
## 17 11.05315 0.1501583 9.066678  
## 18 11.05264 0.1514041 9.065762  
## 19 11.05514 0.1497493 9.068597  
## 20 11.05826 0.1479154 9.073454  
##   
## RMSE was used to select the optimal model using the smallest value.  
## The final value used for the model was mtry = 18.

## Saving 7 x 7 in image  
## Saving 7 x 7 in image



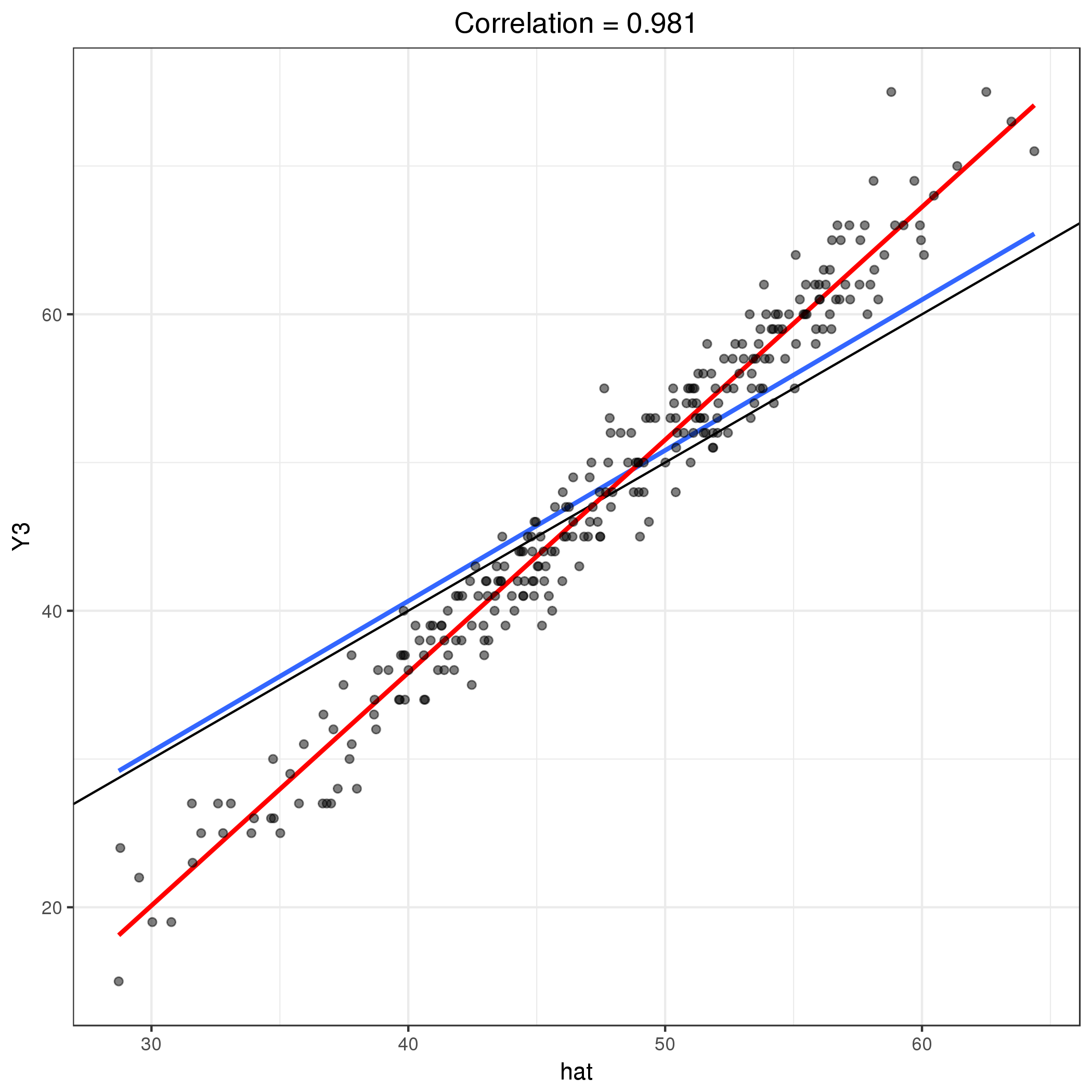
figures/Y3Training.png

##   
## Call:  
## randomForest(x = x, y = y, mtry = param$mtry, localImp = TRUE, nthreads = 8)   
## Type of random forest: regression  
## Number of trees: 500  
## No. of variables tried at each split: 18  
##   
## Mean of squared residuals: 121.1754  
## % Var explained: 12.81

## rf variable importance  
##   
## only 20 most important variables shown (out of 163)  
##   
## Overall  
## zipcode97702 100.00  
## SEPTI\_total 95.73  
## childAge 92.36  
## ECBI\_intensity\_T\_score 86.91  
## childEthnicityNot Hispanic/Latino 83.40  
## MAPS\_LC 82.53  
## ECBI\_problem\_raw\_score 73.89  
## SEPTI\_r\_clinical\_cutoff 69.33  
## MAPS\_POS 68.61  
## zipcode97229 68.53  
## MAPS\_NEG 67.96  
## parentMaritalStatusDivorced 67.46  
## parentAge 66.34  
## parentSituationCo-parenting in separate households 65.59  
## SEPTI\_discipline 64.35  
## childRaceAsian 61.26  
## parentRaceAsian 61.08  
## MAPS\_SP 60.78  
## parentSexMale 59.25  
## MAPS\_PR 58.67

## RMSE Rsquared MAE   
## 4.7925309 0.9618845 3.8682296

## Y3 hat  
## Y3 1.0000000 0.9807571  
## hat 0.9807571 1.0000000

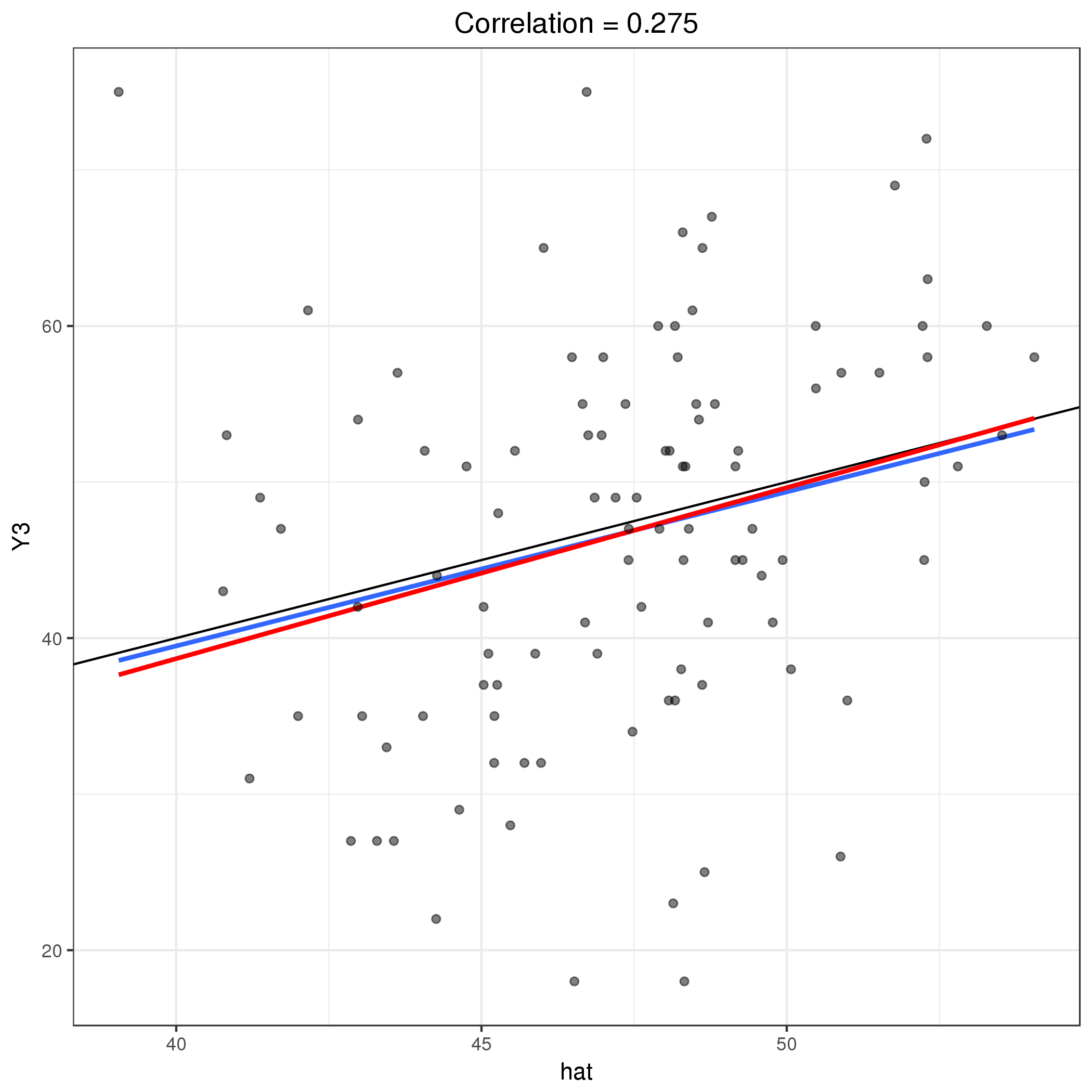


plot of chunk Y3Training-predict

Evaluate model on the validation sample.

## RMSE Rsquared MAE   
## 12.04030176 0.07588566 9.67297644

## Y3 hat  
## Y3 1.0000000 0.2754735  
## hat 0.2754735 1.0000000



plot of chunk Y3Validation-predict