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

Benjamin Chan (chanb@ohsu.edu)

2018-06-27

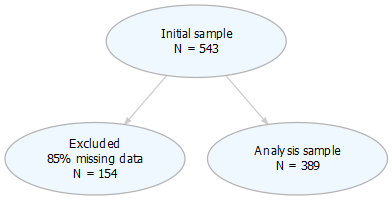
# Import Andrew’s SPSS data

Map new names to variables.

|  |  |
| --- | --- |
| oldnames | newnames |
| record\_id | id |
| eng\_span | languageSurvey |
| children\_totv\_1 | totalChildren |
| oldest\_middle\_youngest | birthOrder |
| child\_sexv\_1 | childSex |
| child\_age\_years | childAge |
| 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 |
| 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 |
| number\_parents | parentsNumber |
| parent\_to\_child\_ratio | parentChildRatio |
| zipcode\_classification\_combined | zipcodeClass |
| 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 | ECBI\_Opp |
| ECBI\_Inatt | ECBI\_Inatt |
| ECBI\_Cond | ECBI\_Cond |
| 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 |
| PCB1\_Total | PCB1\_Total |
| PCB1\_CondEmot | PCB1\_CondEmot |
| PCB1\_DevHab | PCB1\_DevHab |
| PCB2\_Tot | PCB2\_Tot |
| PCB3\_Total | PCB3\_Total |
| PBC3\_PCPonly | PCB3\_PCPonly |
| PCB3\_Person | PCB3\_Person |
| PCB3\_Resource | PCB3\_Resource |

Build analysis data set. Exclude if missing any dependent variable, PCB1\_Total, PCB2\_Tot, PCB3\_Total. Exclude rows if there are a high proportion of row-wise NA.

## PCB1\_Total PCB2\_Tot PCB3\_Total   
## Min. :18.00 Min. : 6.00 Min. :15.0   
## 1st Qu.:58.00 1st Qu.:22.00 1st Qu.:39.0   
## Median :71.00 Median :25.00 Median :48.0   
## Mean :67.89 Mean :24.54 Mean :47.6   
## 3rd Qu.:81.00 3rd Qu.:28.00 3rd Qu.:57.0   
## Max. :90.00 Max. :30.00 Max. :75.0



figures/flowChart.png

# Preprocess data

Initial preprocesssing that needs to be done that is common to PCB1\_Total, PCB2\_Tot, and PCB3\_Total.

p <- 0.75

Split data set into 75:25 training:validation samples.

inTrain <- createDataPartition(df$id, p = p)  
dfTrain <- df[inTrain$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 = 264

nzv <-   
 dfTrain %>%   
 select(-c(id,   
 PCB1\_Total, PCB1\_CondEmot, PCB1\_DevHab,   
 PCB2\_Tot,   
 PCB3\_Total, PCB3\_PCPonly, PCB3\_Person, PCB3\_Resource)) %>%   
 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 | 72.25000 | 0.6825939 | FALSE | TRUE |
| childRaceAfrAm | 19.85714 | 0.6825939 | FALSE | TRUE |
| childRaceAIAN | 35.50000 | 0.6825939 | FALSE | TRUE |
| childRaceNHPI | 47.66667 | 0.6825939 | FALSE | TRUE |
| childRaceOther | 21.46154 | 0.6825939 | FALSE | TRUE |
| parentRaceAfrAm | 40.71429 | 0.6825939 | FALSE | TRUE |
| parentRaceAIAN | 35.50000 | 0.6825939 | FALSE | TRUE |
| parentRaceNHPI | 57.40000 | 0.6825939 | FALSE | TRUE |
| parentRaceOther | 25.54545 | 0.6825939 | FALSE | TRUE |
| internet | 35.50000 | 0.6825939 | FALSE | TRUE |

dfTrainPreProc1 <-  
 dfTrain %>%   
 select(-one\_of(nzv$varname))  
dfOutcomes <-   
 dfTrainPreProc1 %>%   
 select(c(id,  
 PCB1\_Total, PCB1\_CondEmot, PCB1\_DevHab,   
 PCB2\_Tot,   
 PCB3\_Total, PCB3\_PCPonly, PCB3\_Person, PCB3\_Resource))  
dfTrainPreProc2 <-   
 dfTrainPreProc1 %>%   
 select(-c(id,  
 PCB1\_Total, PCB1\_CondEmot, PCB1\_DevHab,   
 PCB2\_Tot,   
 PCB3\_Total, PCB3\_PCPonly, PCB3\_Person, PCB3\_Resource))  
preProc <-  
 dfTrainPreProc2 %>%   
 preProcess(method = c("nzv", "corr", "knnImpute"), verbose = TRUE)

## 2 highly correlated predictors were removed.  
## Calculating 32 means for centering  
## Calculating 32 standard deviations for scaling

preProc

## Created from 264 samples and 54 variables  
##   
## Pre-processing:  
## - centered (32)  
## - ignored (20)  
## - 5 nearest neighbor imputation (32)  
## - removed (2)  
## - scaled (32)

dfTrainPreProc3 <- predict(preProc, dfTrainPreProc2)  
dfTrainPreProc <- bind\_cols(dfOutcomes, dfTrainPreProc3)  
message(sprintf("Number of complete cases after imputation = %d",  
 complete.cases(dfTrainPreProc) %>% sum()))

## Number of complete cases after imputation = 291

save(dfTrainPreProc, dfValid, dfTrain, nzv, preProc, df, file = "data/processed/dataframes.RData")  
rm(dfTrainPreProc1, dfTrainPreProc2, dfTrainPreProc3)

Set the control parameters.

ctrl <- trainControl(method = "repeatedcv",  
 number = 10,  
 repeats = 10,  
 savePredictions = TRUE,  
 allowParallel = TRUE,  
 search = "random")  
cores <- 24

Set the model and tuning parameter grid.

library(earth)

## Loading required package: plotmo

## Loading required package: plotrix

## Loading required package: TeachingDemos

citation("earth")

##   
## To cite package 'earth' in publications use:  
##   
## Stephen Milborrow. Derived from mda:mars by Trevor Hastie and  
## Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas  
## Lumley's leaps wrapper. (2018). earth: Multivariate Adaptive  
## Regression Splines. R package version 4.6.2.  
## https://CRAN.R-project.org/package=earth  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {earth: Multivariate Adaptive Regression Splines},  
## author = {Stephen Milborrow. Derived from mda:mars by Trevor Hastie and Rob Tibshirani. Uses Alan Miller's Fortran utilities with Thomas Lumley's leaps wrapper.},  
## year = {2018},  
## note = {R package version 4.6.2},  
## url = {https://CRAN.R-project.org/package=earth},  
## }  
##   
## ATTENTION: This citation information has been auto-generated from  
## the package DESCRIPTION file and may need manual editing, see  
## 'help("citation")'.

method <- "bagEarth"  
modelLookup(method) %>% kable()

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | parameter | label | forReg | forClass | probModel |
| bagEarth | nprune | #Terms | TRUE | TRUE | TRUE |
| bagEarth | degree | Product Degree | TRUE | TRUE | TRUE |

grid <- expand.grid(nprune = c(seq(2, 9, 1), seq(10, 20, 5)),  
 degree = 1)  
grid %>% kable()

|  |  |
| --- | --- |
| nprune | degree |
| 2 | 1 |
| 3 | 1 |
| 4 | 1 |
| 5 | 1 |
| 6 | 1 |
| 7 | 1 |
| 8 | 1 |
| 9 | 1 |
| 10 | 1 |
| 15 | 1 |
| 20 | 1 |

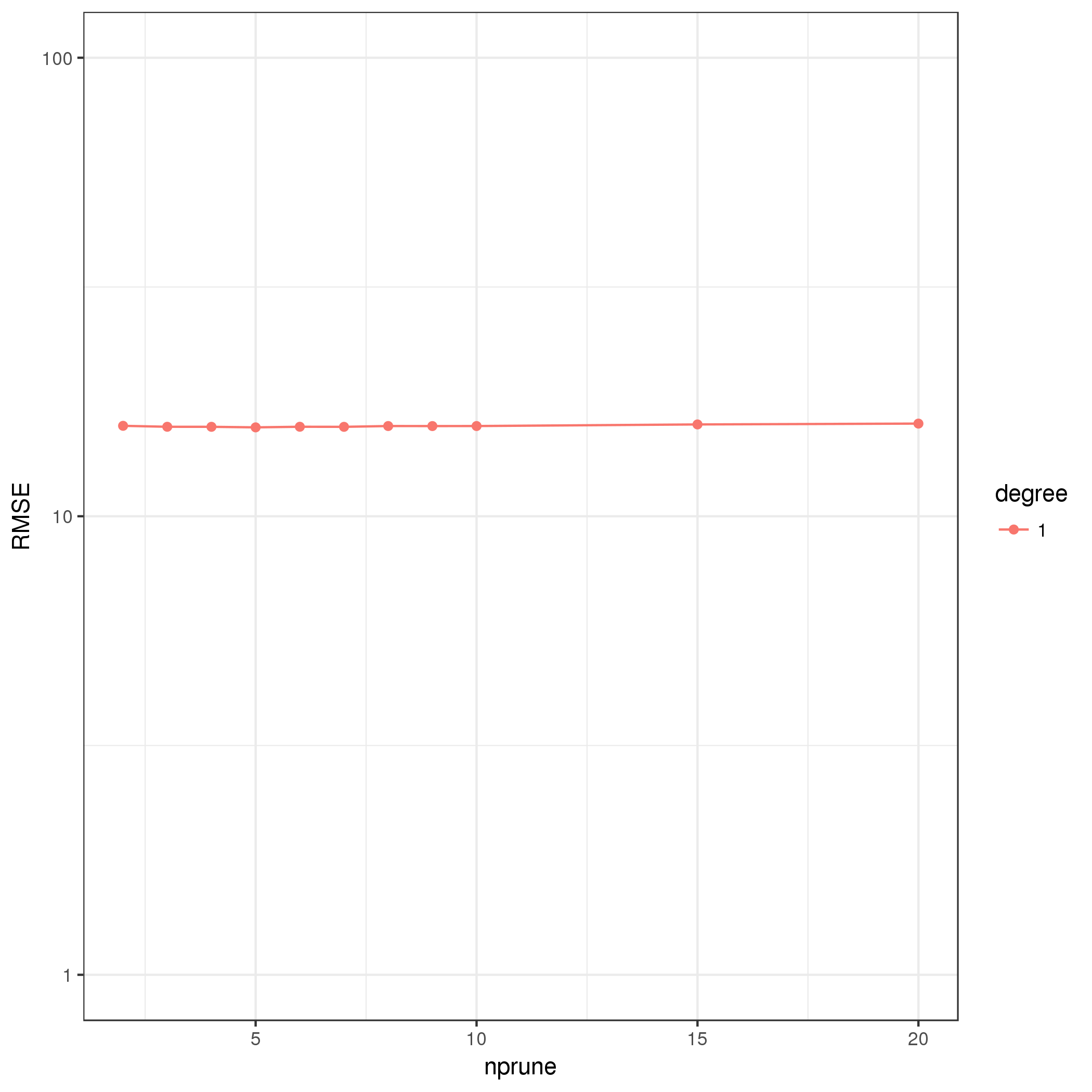
# Model PCB1

## PCB1 Total

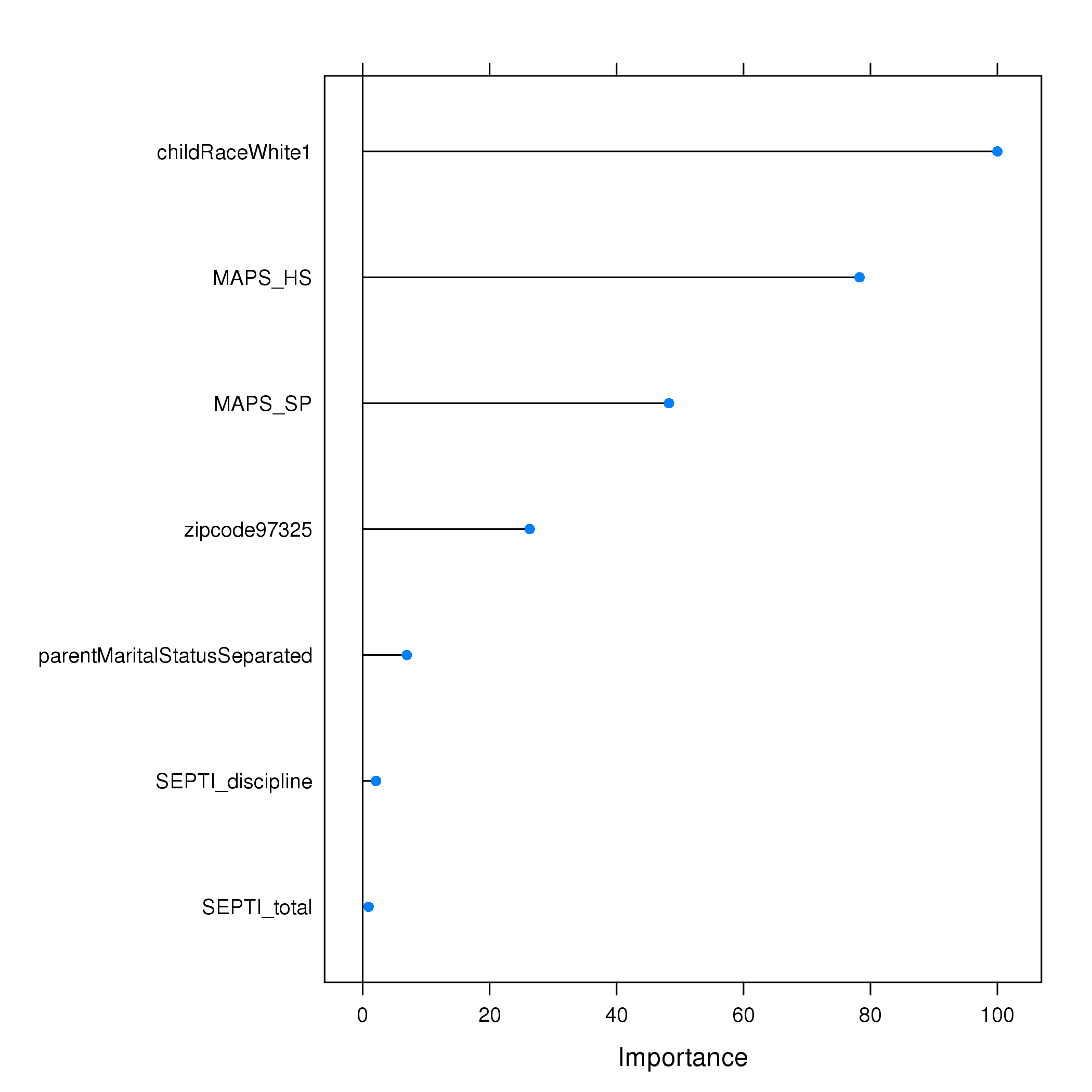
Prediction model for PCB1.

Train model over the tuning parameters.

## Bagged MARS   
##   
## 293 samples  
## 52 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 261, 263, 263, 261, 261, 262, ...   
## Resampling results across tuning parameters:  
##   
## nprune RMSE Rsquared MAE   
## 2 15.74837 0.04075392 12.87834  
## 3 15.67317 0.04823170 12.77305  
## 4 15.66999 0.05079536 12.76927  
## 5 15.62701 0.05904268 12.73300  
## 6 15.67466 0.05568217 12.77803  
## 7 15.66848 0.06020719 12.73589  
## 8 15.73151 0.05595855 12.80302  
## 9 15.72864 0.05860294 12.77428  
## 10 15.73144 0.06245847 12.75535  
## 15 15.85905 0.06046205 12.79040  
## 20 15.92524 0.06566826 12.79684  
##   
## Tuning parameter 'degree' was held constant at a value of 1  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were nprune = 5 and degree = 1.



plot of chunk PCB1\_Total\_Training

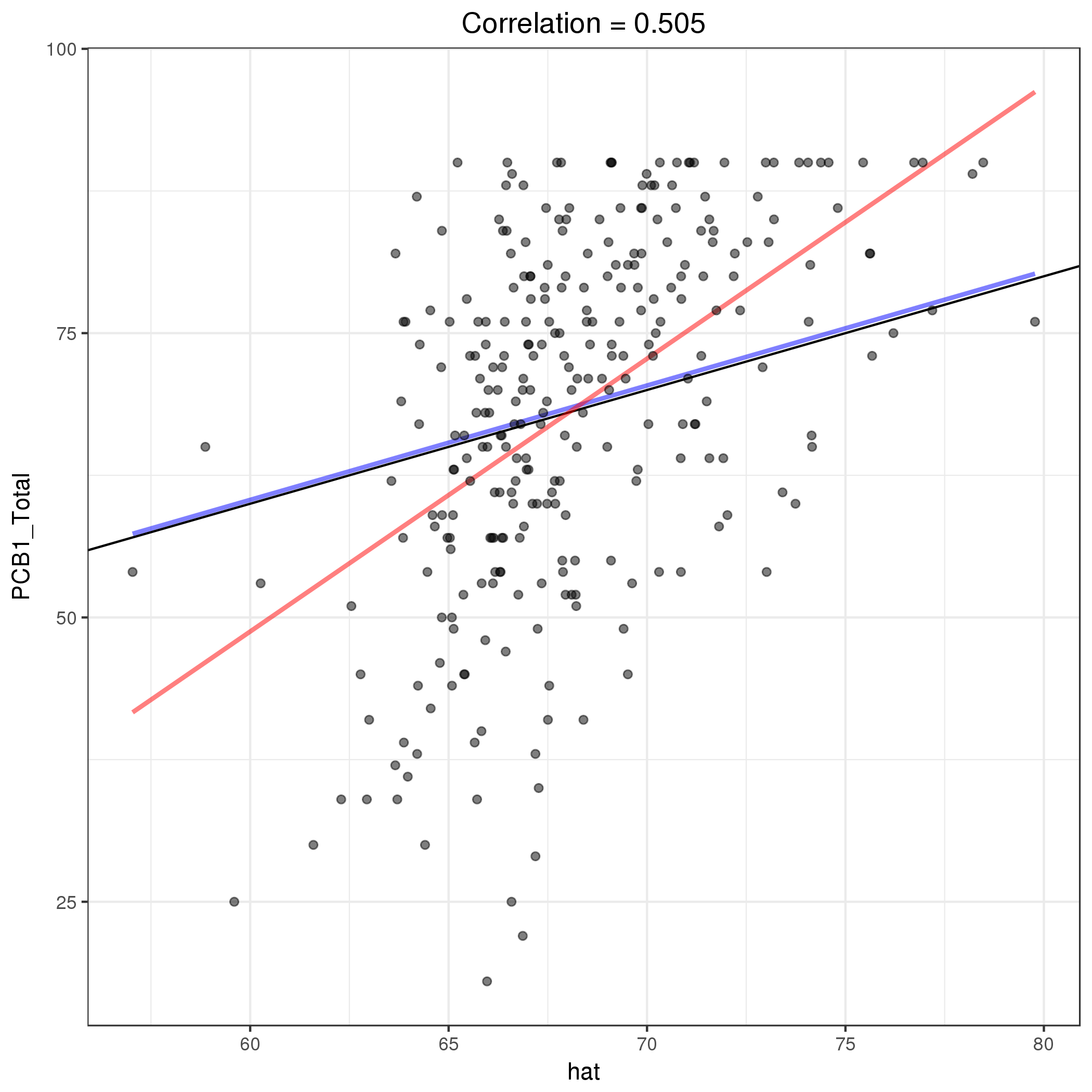


plot of chunk PCB1\_Total\_Training-varImp

|  |  |
| --- | --- |
| variable | Overall |
| childRaceWhite1 | 100.000000 |
| MAPS\_HS | 78.278812 |
| MAPS\_SP | 48.261261 |
| zipcode97325 | 26.312871 |
| parentMaritalStatusSeparated | 6.953803 |
| SEPTI\_discipline | 2.109540 |
| SEPTI\_total | 0.937654 |

## RMSE Rsquared MAE   
## 14.4209179 0.2551321 11.7697091

## PCB1\_Total hat  
## PCB1\_Total 1.000000 0.505106  
## hat 0.505106 1.000000

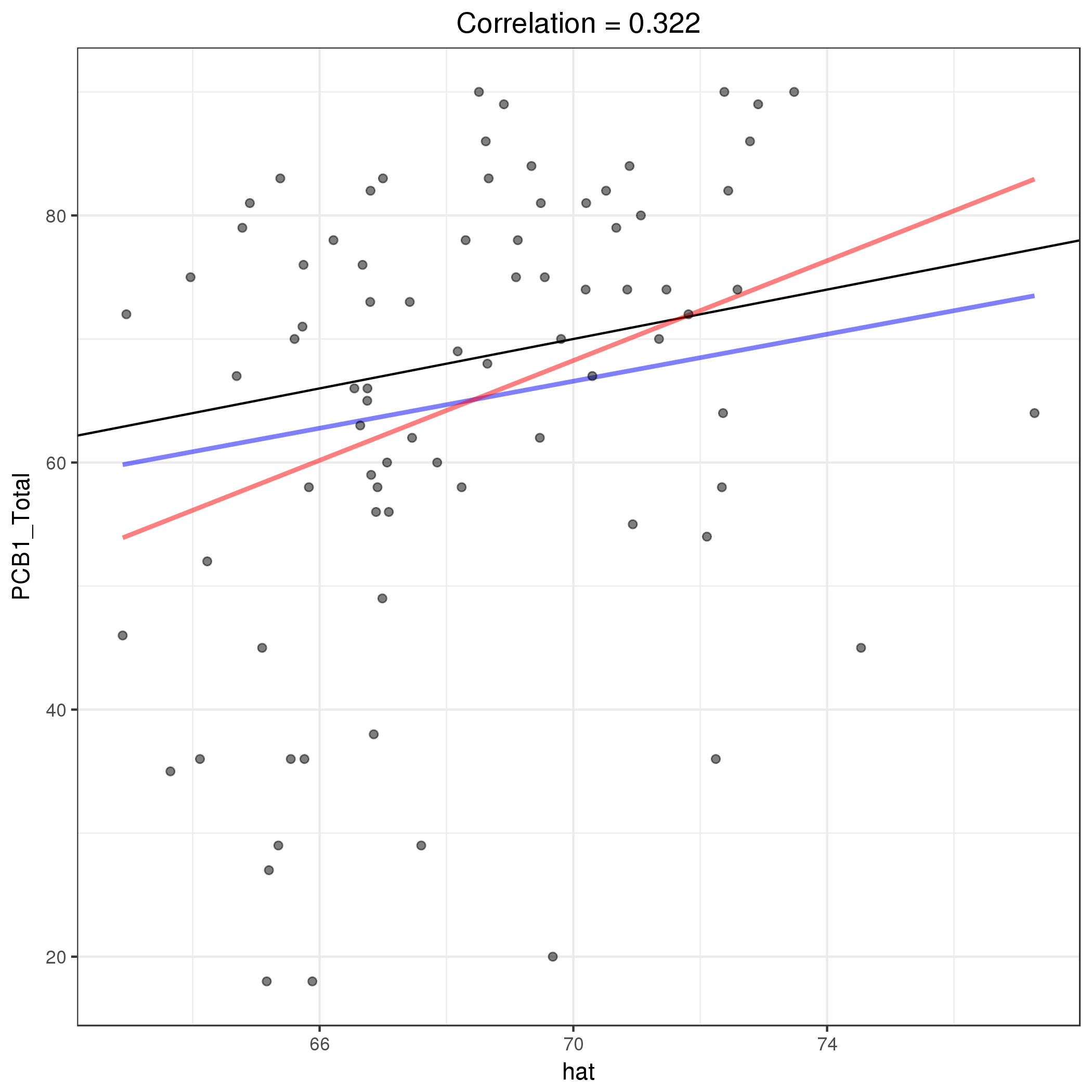


plot of chunk PCB1\_Total\_Training-predict

Evaluate model on the validation sample.

## RMSE Rsquared MAE   
## 17.9851222 0.1039878 13.8753277

## PCB1\_Total hat  
## PCB1\_Total 1.0000000 0.3224714  
## hat 0.3224714 1.0000000



plot of chunk PCB1\_Total\_Validation-predict

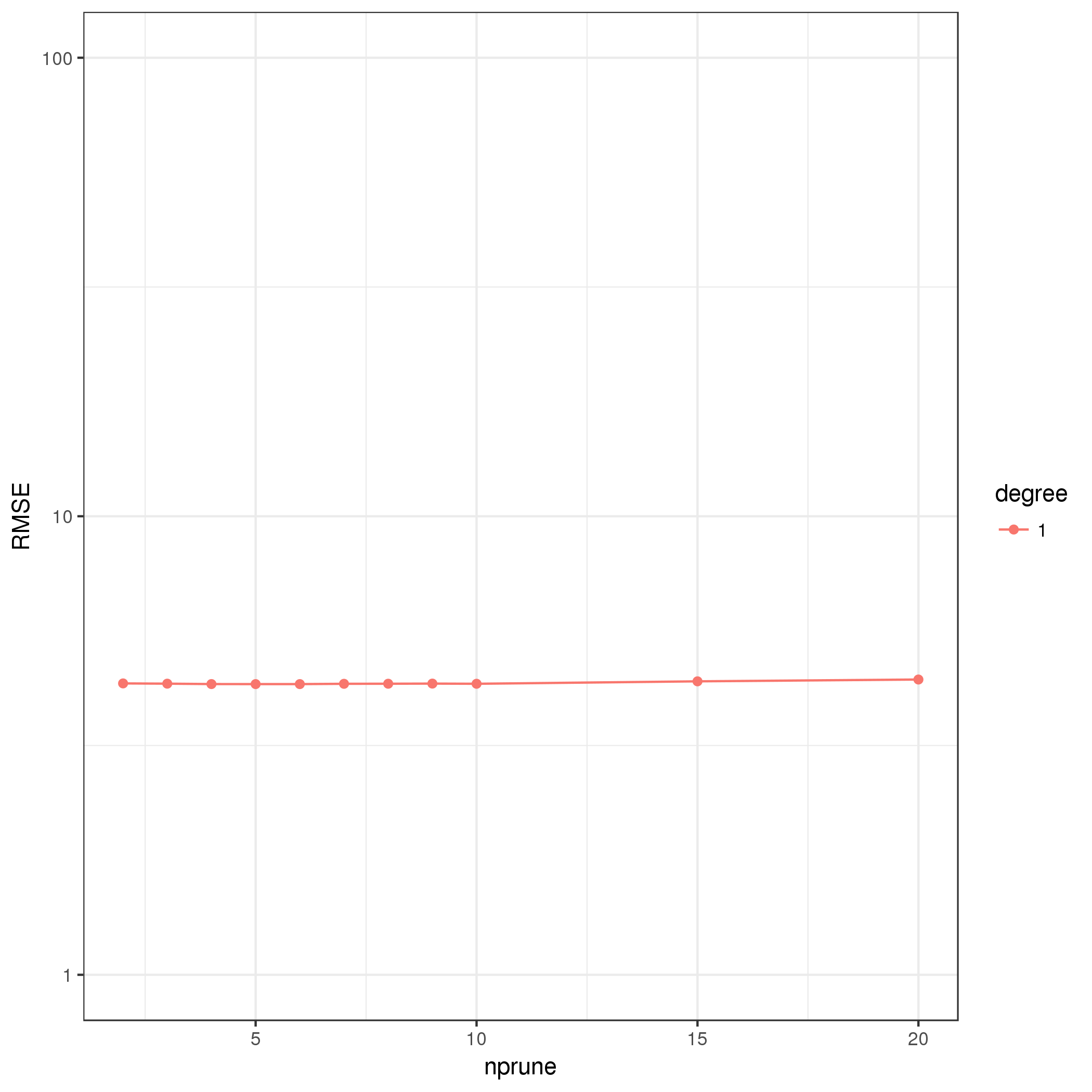
# Model PCB2

## Total PCB2

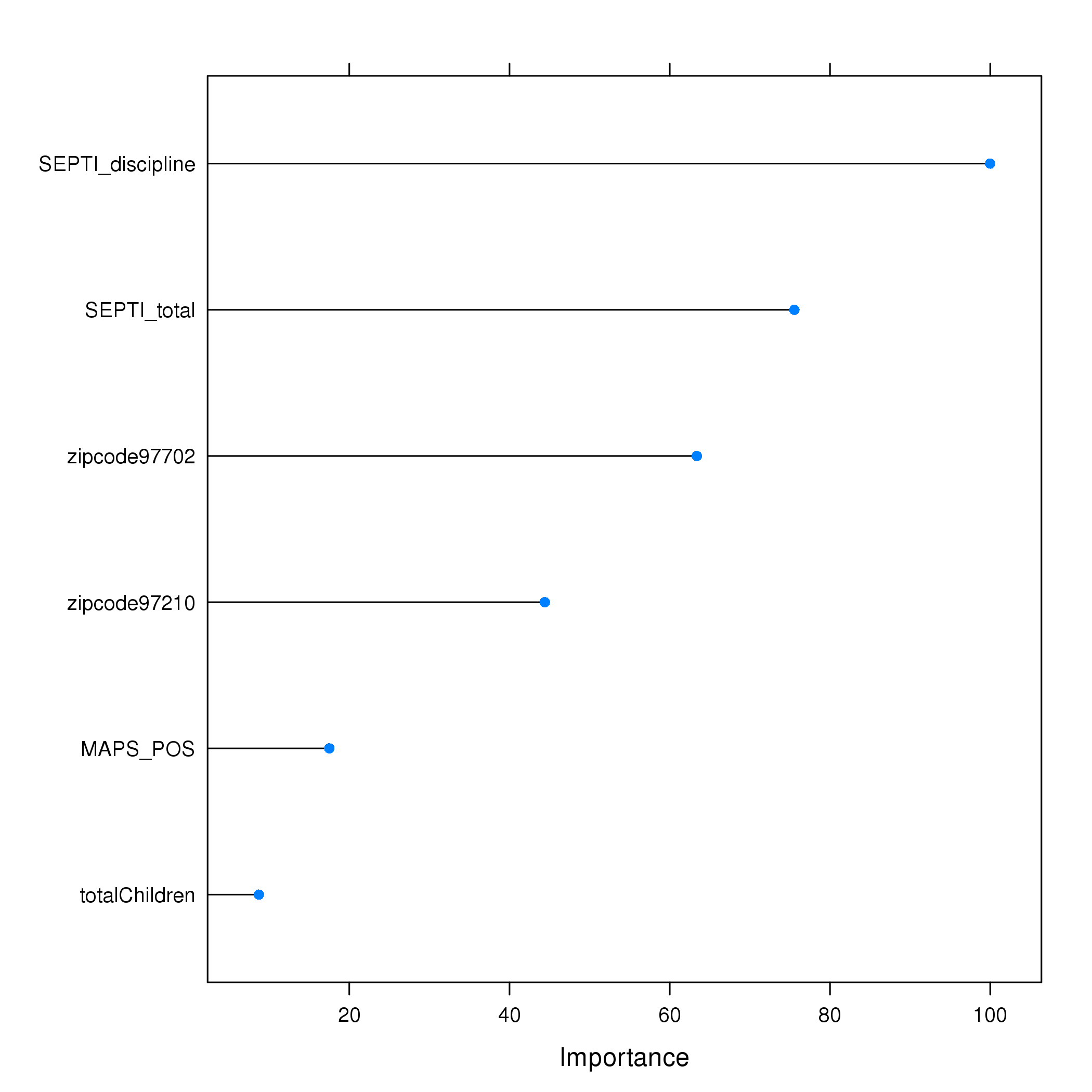
Prediction model for PCB2\_Total.

Train model over the tuning parameters.

## Bagged MARS   
##   
## 293 samples  
## 52 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 263, 261, 263, 262, 261, 262, ...   
## Resampling results across tuning parameters:  
##   
## nprune RMSE Rsquared MAE   
## 2 4.319944 0.05273309 3.294298  
## 3 4.314337 0.05819237 3.284545  
## 4 4.305675 0.05651808 3.275975  
## 5 4.305110 0.05401634 3.276022  
## 6 4.305036 0.05800955 3.267395  
## 7 4.311454 0.05720017 3.273882  
## 8 4.313036 0.05651665 3.274898  
## 9 4.315909 0.05702546 3.278187  
## 10 4.312015 0.06007262 3.275315  
## 15 4.365279 0.05661681 3.309819  
## 20 4.405090 0.05862837 3.339554  
##   
## Tuning parameter 'degree' was held constant at a value of 1  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were nprune = 6 and degree = 1.



plot of chunk PCB2\_Tot\_Training

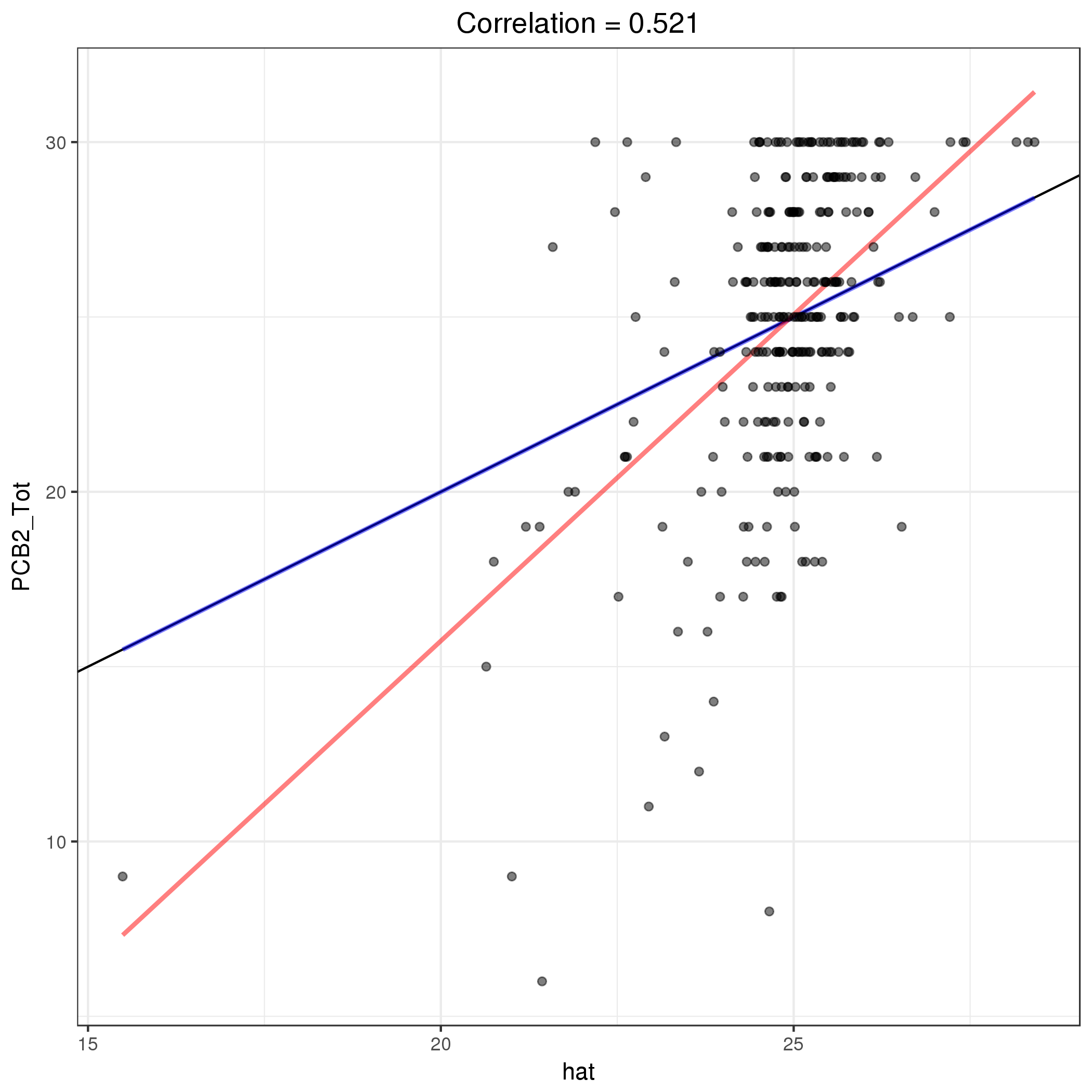


plot of chunk PCB2\_Total\_Training-varImp

|  |  |
| --- | --- |
| variable | Overall |
| SEPTI\_discipline | 100.000000 |
| SEPTI\_total | 75.556611 |
| zipcode97702 | 63.379081 |
| zipcode97210 | 44.409026 |
| MAPS\_POS | 17.512654 |
| totalChildren | 8.708317 |

## RMSE Rsquared MAE   
## 3.8654487 0.2714033 2.9648018

## PCB2\_Tot hat  
## PCB2\_Tot 1.0000000 0.5209638  
## hat 0.5209638 1.0000000

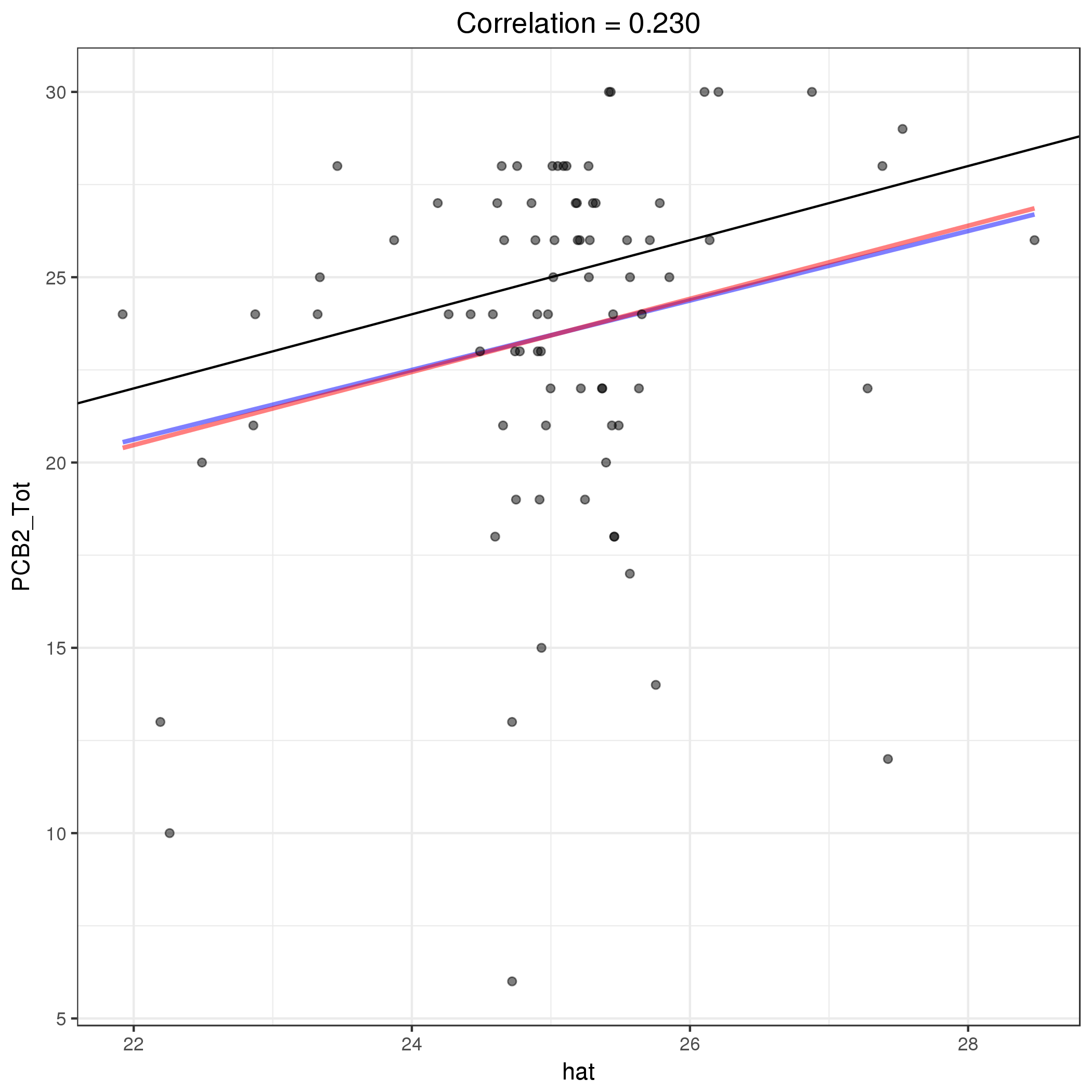


plot of chunk PCB2\_Tot\_Training-predict

Evaluate model on the validation sample.

## RMSE Rsquared MAE   
## 4.93520927 0.05291807 3.48080867

## PCB2\_Tot hat  
## PCB2\_Tot 1.0000000 0.2300393  
## hat 0.2300393 1.0000000



plot of chunk PCB2\_Tot\_Validation-predict

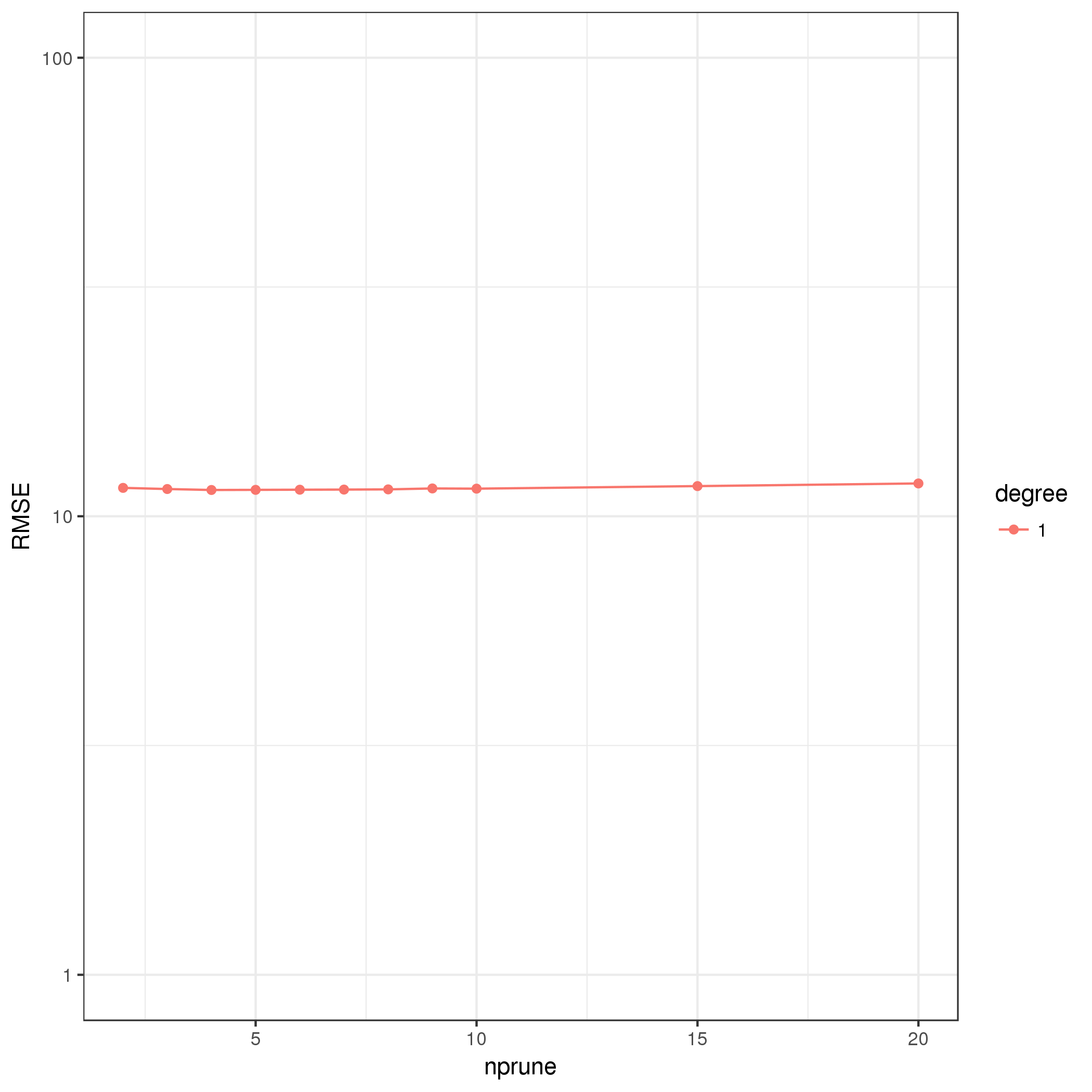
# Model PCB3

## Total PCB3

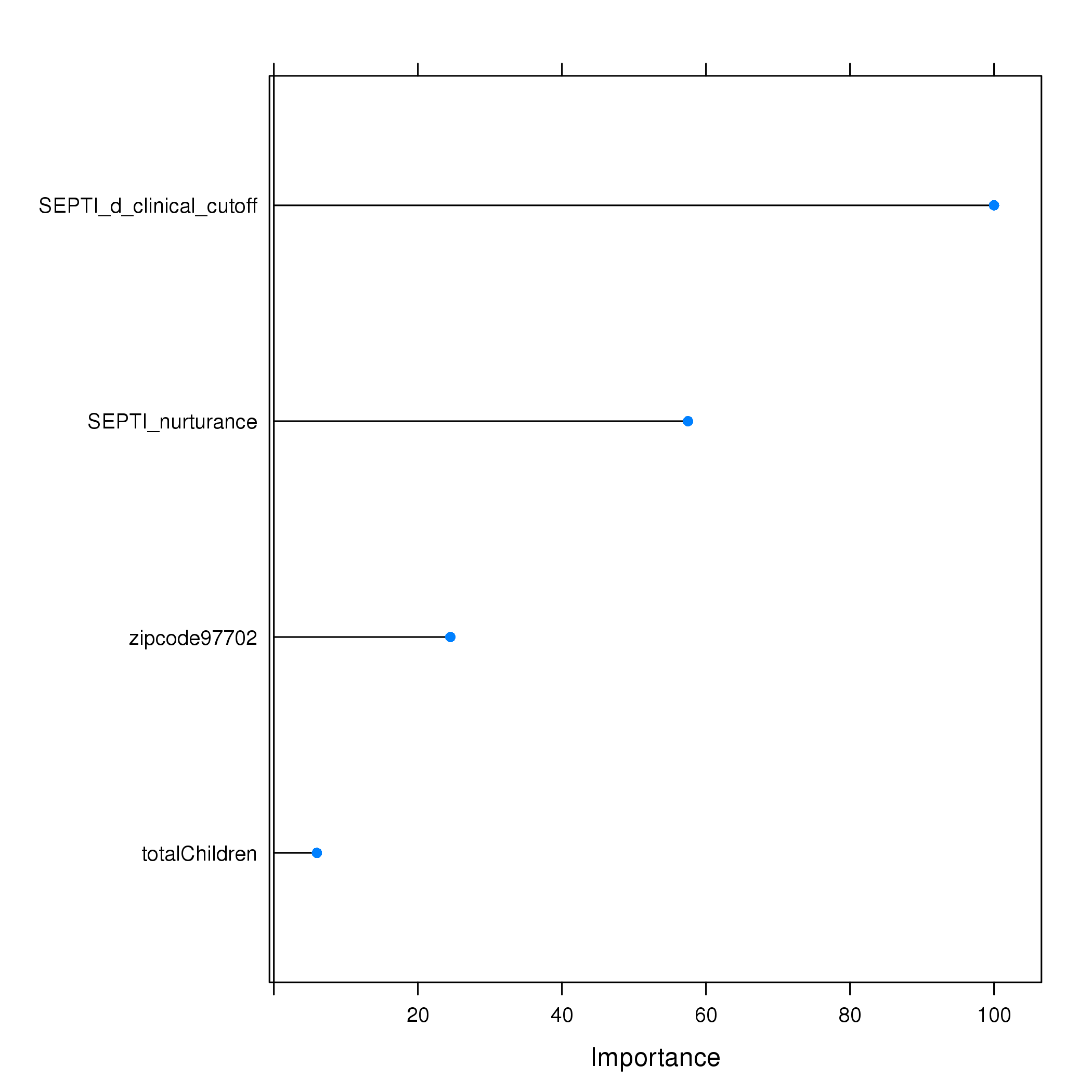
Prediction model for PCB3\_Total.

Train model over the tuning parameters.

## Bagged MARS   
##   
## 293 samples  
## 52 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (10 fold, repeated 10 times)   
## Summary of sample sizes: 262, 262, 262, 263, 261, 261, ...   
## Resampling results across tuning parameters:  
##   
## nprune RMSE Rsquared MAE   
## 2 11.53176 0.09508904 9.570356  
## 3 11.46565 0.10178068 9.500293  
## 4 11.41050 0.11083703 9.437381  
## 5 11.41801 0.10699896 9.425560  
## 6 11.42676 0.10372935 9.411713  
## 7 11.43381 0.10551998 9.404470  
## 8 11.44510 0.10373924 9.400230  
## 9 11.49922 0.10271975 9.425596  
## 10 11.48670 0.10735322 9.406348  
## 15 11.63536 0.10202053 9.500001  
## 20 11.79091 0.09035788 9.600144  
##   
## Tuning parameter 'degree' was held constant at a value of 1  
## RMSE was used to select the optimal model using the smallest value.  
## The final values used for the model were nprune = 4 and degree = 1.



plot of chunk PCB3\_Total\_Training

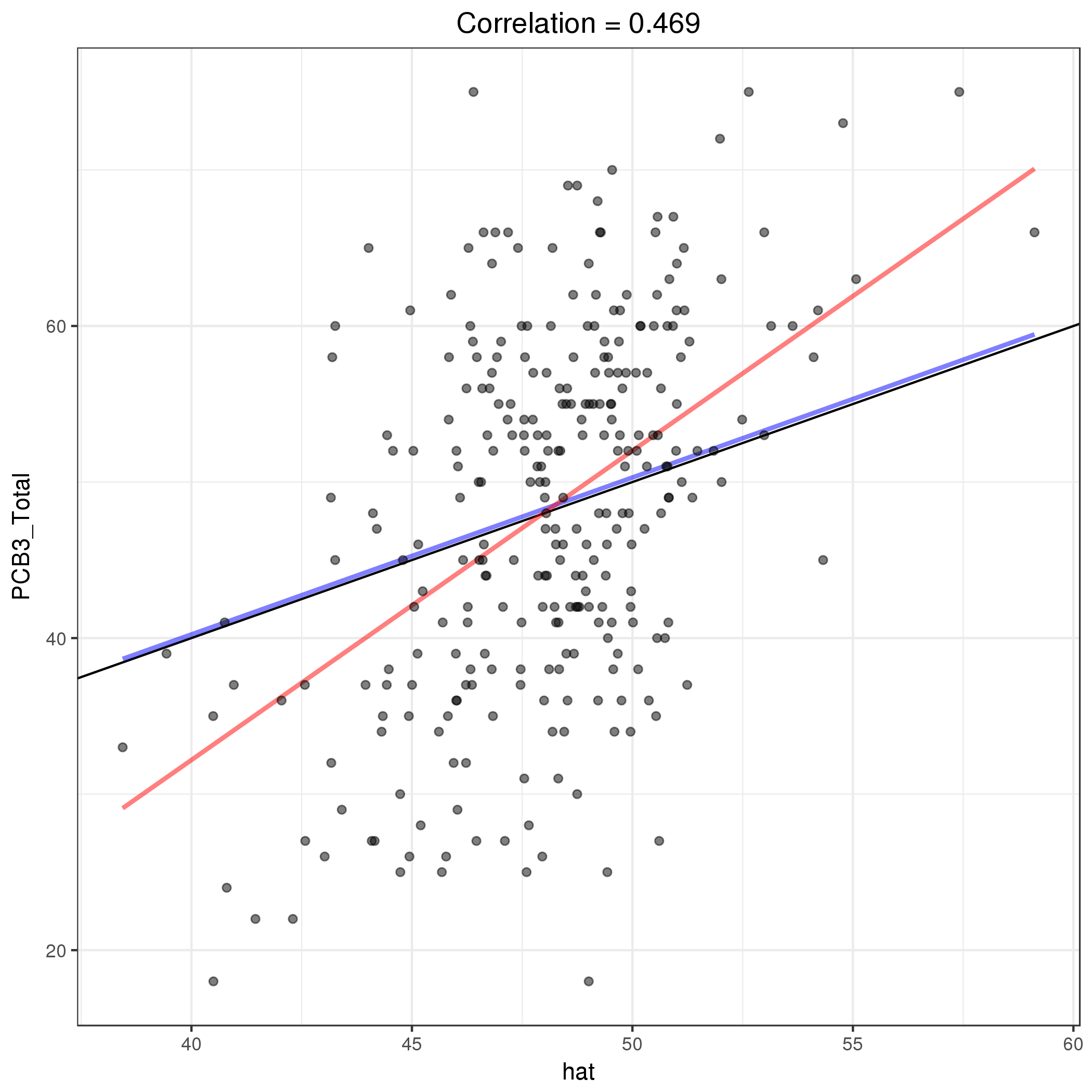


plot of chunk PCB3\_Total\_Training-varImp

|  |  |
| --- | --- |
| variable | Overall |
| SEPTI\_d\_clinical\_cutoff | 100.000000 |
| SEPTI\_nurturance | 57.507706 |
| zipcode97702 | 24.510919 |
| totalChildren | 5.975344 |

## RMSE Rsquared MAE   
## 10.804803 0.220301 8.957826

## PCB3\_Total hat  
## PCB3\_Total 1.0000000 0.4693624  
## hat 0.4693624 1.0000000

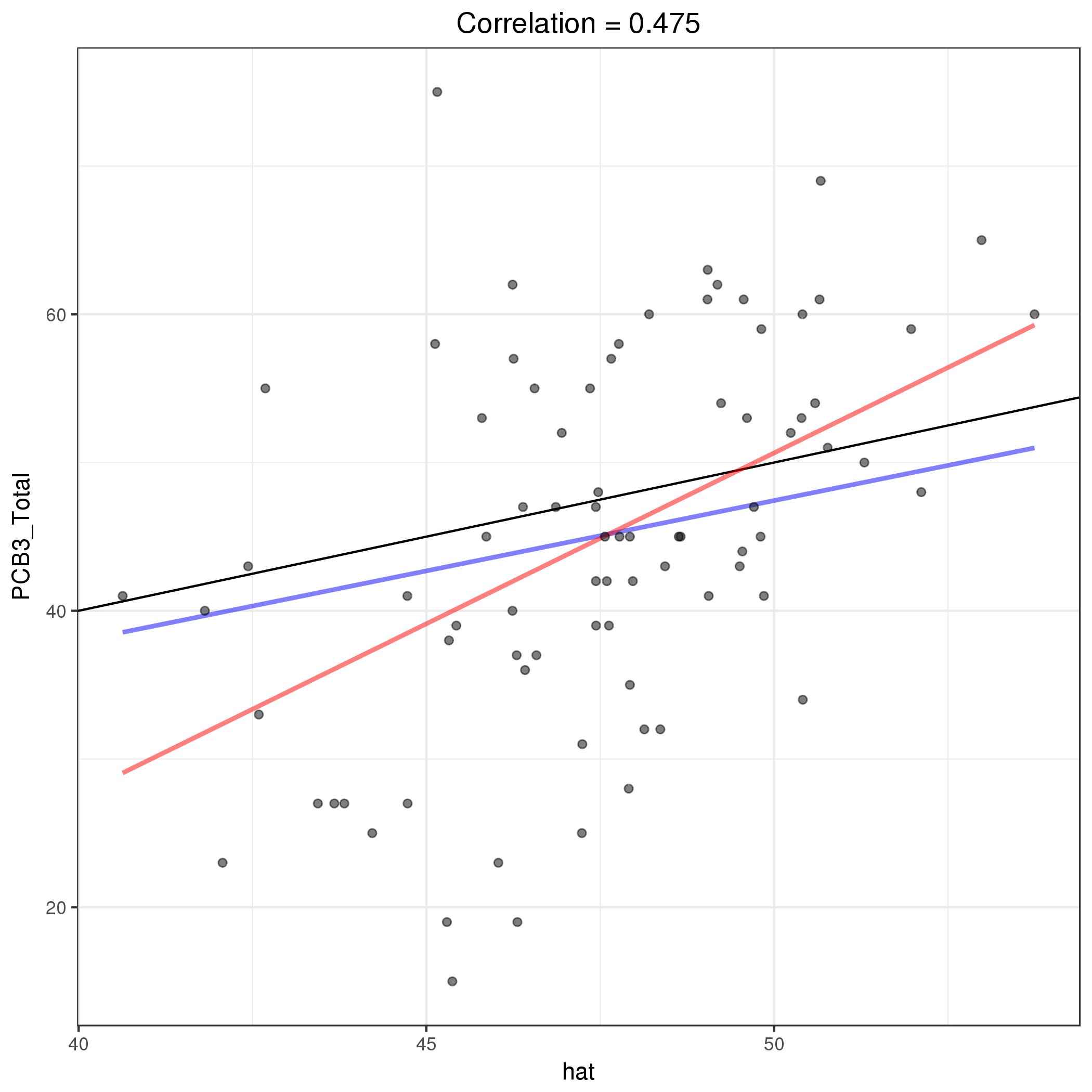


plot of chunk PCB3\_Total\_Training-predict

Evaluate model on the validation sample.

## RMSE Rsquared MAE   
## 11.9969623 0.2258918 9.6100738

## PCB3\_Total hat  
## PCB3\_Total 1.0000000 0.4752808  
## hat 0.4752808 1.0000000



plot of chunk PCB3\_TotalValidation-predict