Week 11 Report

Jackson Tucker & Dylan Denner

4/6/2021

## Exercises

1. We will now perform cross-validation on a simulated data set.
2. Generate a simulated data set as follows:

set.seed(1) x=rnorm(100) y=x-2\*x^2+rnorm (100)

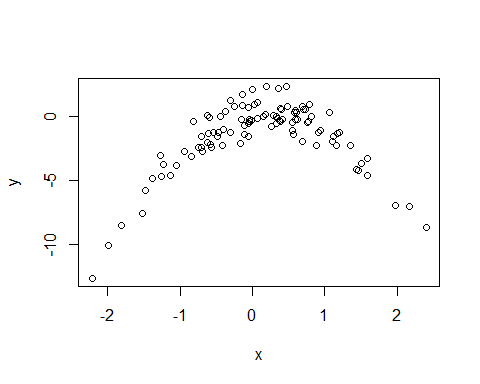
In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

library(boot)  
setwd("C:/Users/dylan/Documents/R/MATH5530/Week11")  
  
set.seed(1)  
x <- rnorm(100)  
y <- x -2\*x^2 + rnorm(100)  
  
df <- data.frame(x,y)

Here, n = 100, p = 2

1. Create a scatterplot of X against Y . Comment on what you find.

plot(x,y)



There is a very clear nonlinear, quadratic relationship between x and y.

1. Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:
2. Y = β0 + β1X + e
3. Y = β0 + β1X + β2X2 + e
4. Y = β0 + β1X + β2X2 + β3X3 + e
5. Y = β0 + β1X + β2X2 + β3X3 + β4X4 + e

set.seed(100)  
  
model1 <- glm(y~x, data = df)  
LOOCVerror1 <- cv.glm(df,model1)$delta  
LOOCVerror1

## [1] 7.288162 7.284744

model2 <- glm(y~poly(x,2), data = df)  
LOOCVerror2 <- cv.glm(df,model2)$delta  
LOOCVerror2

## [1] 0.9374236 0.9371789

model3 <- glm(y~poly(x,3), data = df)  
LOOCVerror3 <- cv.glm(df,model3)$delta  
LOOCVerror3

## [1] 0.9566218 0.9562538

model4 <- glm(y~poly(x,4), data = df)  
LOOCVerror4 <- cv.glm(df,model4)$delta  
LOOCVerror4

## [1] 0.9539049 0.9534453

1. Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

set.seed(200)  
x <- rnorm(100)  
y <- x -2\*x^2 + rnorm(100)  
  
df <- data.frame(x,y)  
  
model1\_2 <- glm(y~x, data = df)  
LOOCVerror1\_2 <- cv.glm(df,model1\_2)$delta  
LOOCVerror1\_2

## [1] 8.750477 8.744707

model2\_2 <- glm(y~poly(x,2), data = df)  
LOOCVerror2\_2 <- cv.glm(df,model2\_2)$delta  
LOOCVerror2\_2

## [1] 0.7311288 0.7308963

model3\_2 <- glm(y~poly(x,3), data = df)  
LOOCVerror3\_2 <- cv.glm(df,model3\_2)$delta  
LOOCVerror3\_2

## [1] 0.8494860 0.8482649

model4\_2 <- glm(y~poly(x,4), data = df)  
LOOCVerror4\_2 <- cv.glm(df,model4\_2)$delta  
LOOCVerror4\_2

## [1] 0.8876290 0.8858296

1. Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.
2. Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

summary(model1)

##   
## Call:  
## glm(formula = y ~ x, data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -9.5161 -0.6800 0.6812 1.5491 3.8183   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.6254 0.2619 -6.205 1.31e-08 \*\*\*  
## x 0.6925 0.2909 2.380 0.0192 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 6.760719)  
##   
## Null deviance: 700.85 on 99 degrees of freedom  
## Residual deviance: 662.55 on 98 degrees of freedom  
## AIC: 478.88  
##   
## Number of Fisher Scoring iterations: 2

summary(model2)

##   
## Call:  
## glm(formula = y ~ poly(x, 2), data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9650 -0.6254 -0.1288 0.5803 2.2700   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.5500 0.0958 -16.18 < 2e-16 \*\*\*  
## poly(x, 2)1 6.1888 0.9580 6.46 4.18e-09 \*\*\*  
## poly(x, 2)2 -23.9483 0.9580 -25.00 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.9178258)  
##   
## Null deviance: 700.852 on 99 degrees of freedom  
## Residual deviance: 89.029 on 97 degrees of freedom  
## AIC: 280.17  
##   
## Number of Fisher Scoring iterations: 2

summary(model3)

##   
## Call:  
## glm(formula = y ~ poly(x, 3), data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.9765 -0.6302 -0.1227 0.5545 2.2843   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.55002 0.09626 -16.102 < 2e-16 \*\*\*  
## poly(x, 3)1 6.18883 0.96263 6.429 4.97e-09 \*\*\*  
## poly(x, 3)2 -23.94830 0.96263 -24.878 < 2e-16 \*\*\*  
## poly(x, 3)3 0.26411 0.96263 0.274 0.784   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.9266599)  
##   
## Null deviance: 700.852 on 99 degrees of freedom  
## Residual deviance: 88.959 on 96 degrees of freedom  
## AIC: 282.09  
##   
## Number of Fisher Scoring iterations: 2

summary(model4)

##   
## Call:  
## glm(formula = y ~ poly(x, 4), data = df)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.0550 -0.6212 -0.1567 0.5952 2.2267   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -1.55002 0.09591 -16.162 < 2e-16 \*\*\*  
## poly(x, 4)1 6.18883 0.95905 6.453 4.59e-09 \*\*\*  
## poly(x, 4)2 -23.94830 0.95905 -24.971 < 2e-16 \*\*\*  
## poly(x, 4)3 0.26411 0.95905 0.275 0.784   
## poly(x, 4)4 1.25710 0.95905 1.311 0.193   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.9197797)  
##   
## Null deviance: 700.852 on 99 degrees of freedom  
## Residual deviance: 87.379 on 95 degrees of freedom  
## AIC: 282.3  
##   
## Number of Fisher Scoring iterations: 2

## Preperation for Presentation

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3 v purrr 0.3.4  
## v tibble 3.0.5 v stringr 1.4.0  
## v tidyr 1.1.3 v forcats 0.5.1  
## v readr 1.4.0

## -- Conflicts ------------------------------------------ tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

##   
## Attaching package: 'gridExtra'

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

##   
## Attaching package: 'MASS'

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

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:purrr':  
##   
## some

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

## The following object is masked from 'package:boot':  
##   
## logit

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

library(tidycensus)  
  
poverty\_by\_sex\_edu\_attainment <- get\_acs(  
 geography = "school district (unified)",   
 variables = c(total\_population = "B17003\_001",  
 below\_poverty\_num = "B17003\_002",  
 below\_poverty\_num\_men = "B17003\_003",  
 M\_below\_poverty\_less\_than\_HS = "B17003\_004",  
 M\_below\_poverty\_HS = "B17003\_005",  
 M\_below\_poverty\_some\_college = "B17003\_006",  
 M\_below\_poverty\_bachelor\_plus = "B17003\_007",  
 below\_poverty\_num\_women = "B17003\_008",  
 W\_below\_poverty\_less\_than\_HS = "B17003\_009",  
 W\_below\_poverty\_HS = "B17003\_010",  
 W\_below\_poverty\_some\_college = "B17003\_011",  
 W\_below\_poverty\_bachelor\_plus = "B17003\_012",  
 above\_poverty\_num = "B17003\_013",  
 above\_poverty\_num\_men = "B17003\_014",  
 M\_above\_poverty\_less\_than\_HS = "B17003\_015",  
 M\_above\_poverty\_HS = "B17003\_016",  
 M\_above\_poverty\_some\_college = "B17003\_017",  
 M\_above\_poverty\_bachelor\_plus = "B17003\_018",  
 above\_poverty\_num\_women = "B17003\_019",  
 W\_above\_poverty\_less\_than\_HS = "B17003\_020",  
 W\_above\_poverty\_HS = "B17003\_021",  
 W\_above\_poverty\_some\_college = "B17003\_022",  
 W\_above\_poverty\_bachelor\_plus = "B17003\_023"),  
 state = "OH",   
 year = 2019)

## Getting data from the 2015-2019 5-year ACS

head(poverty\_by\_sex\_edu\_attainment)

## # A tibble: 6 x 5  
## GEOID NAME variable estimate moe  
## <chr> <chr> <chr> <dbl> <dbl>  
## 1 3900094 Monroe Local School District,~ total\_population 9268 352  
## 2 3900094 Monroe Local School District,~ below\_poverty\_num 230 133  
## 3 3900094 Monroe Local School District,~ below\_poverty\_num\_men 60 69  
## 4 3900094 Monroe Local School District,~ M\_below\_poverty\_less\_th~ 10 15  
## 5 3900094 Monroe Local School District,~ M\_below\_poverty\_HS 50 67  
## 6 3900094 Monroe Local School District,~ M\_below\_poverty\_some\_co~ 0 18

dim(poverty\_by\_sex\_edu\_attainment)

## [1] 14099 5

length(unique(poverty\_by\_sex\_edu\_attainment$NAME))

## [1] 613

poverty\_by\_sex\_edu\_attainment <- poverty\_by\_sex\_edu\_attainment[,-c(5)]

poverty\_by\_sex\_edu\_attainment <- spread(poverty\_by\_sex\_edu\_attainment, variable, estimate)  
  
head(poverty\_by\_sex\_edu\_attainment)

## # A tibble: 6 x 25  
## GEOID NAME above\_poverty\_n~ above\_poverty\_nu~ above\_poverty\_nu~  
## <chr> <chr> <dbl> <dbl> <dbl>  
## 1 39000~ Monroe Local Scho~ 9038 4420 4618  
## 2 39005~ Manchester Local ~ 2612 1344 1268  
## 3 39043~ Akron City School~ 101084 48564 52520  
## 4 39043~ Alliance City Sch~ 10683 5201 5482  
## 5 39043~ Ashland City Scho~ 14834 7308 7526  
## 6 39043~ Ashtabula Area Ci~ 14633 6820 7813  
## # ... with 20 more variables: below\_poverty\_num <dbl>,  
## # below\_poverty\_num\_men <dbl>, below\_poverty\_num\_women <dbl>,  
## # M\_above\_poverty\_bachelor\_plus <dbl>, M\_above\_poverty\_HS <dbl>,  
## # M\_above\_poverty\_less\_than\_HS <dbl>, M\_above\_poverty\_some\_college <dbl>,  
## # M\_below\_poverty\_bachelor\_plus <dbl>, M\_below\_poverty\_HS <dbl>,  
## # M\_below\_poverty\_less\_than\_HS <dbl>, M\_below\_poverty\_some\_college <dbl>,  
## # total\_population <dbl>, W\_above\_poverty\_bachelor\_plus <dbl>,  
## # W\_above\_poverty\_HS <dbl>, W\_above\_poverty\_less\_than\_HS <dbl>,  
## # W\_above\_poverty\_some\_college <dbl>, W\_below\_poverty\_bachelor\_plus <dbl>,  
## # W\_below\_poverty\_HS <dbl>, W\_below\_poverty\_less\_than\_HS <dbl>,  
## # W\_below\_poverty\_some\_college <dbl>

dim(poverty\_by\_sex\_edu\_attainment)

## [1] 613 25

mod\_poverty\_by\_sex\_edu\_attainment <- poverty\_by\_sex\_edu\_attainment %>% mutate(  
 "tot\_less\_than\_HS" = W\_below\_poverty\_less\_than\_HS + M\_below\_poverty\_less\_than\_HS +   
 W\_above\_poverty\_less\_than\_HS + M\_above\_poverty\_less\_than\_HS,  
 "tot\_HS" = W\_below\_poverty\_HS + M\_below\_poverty\_HS +   
 W\_above\_poverty\_HS + M\_above\_poverty\_HS,  
 "tot\_some\_college" = W\_below\_poverty\_some\_college + M\_below\_poverty\_some\_college +   
 W\_above\_poverty\_some\_college + M\_above\_poverty\_some\_college,  
 "tot\_bachelor\_plus" = W\_below\_poverty\_bachelor\_plus + M\_below\_poverty\_bachelor\_plus +   
 W\_above\_poverty\_bachelor\_plus + M\_above\_poverty\_bachelor\_plus,  
 "tot\_HS\_PLUS" = tot\_HS + tot\_some\_college + tot\_bachelor\_plus,  
 "poverty\_percentage" = below\_poverty\_num/total\_population,  
 "HS\_PLUS\_percentage" = tot\_HS\_PLUS/total\_population)

mod\_poverty\_by\_sex\_edu\_attainment <- dplyr::select(  
 mod\_poverty\_by\_sex\_edu\_attainment,  
 NAME,  
 total\_population,  
 below\_poverty\_num,  
 above\_poverty\_num,  
 poverty\_percentage,  
 tot\_less\_than\_HS,  
 tot\_HS,  
 tot\_some\_college,  
 tot\_bachelor\_plus,  
 HS\_PLUS\_percentage  
)  
  
dim(mod\_poverty\_by\_sex\_edu\_attainment)

## [1] 613 10

file\_path <- "../DATA/BUILDING\_OVERVIEW\_1819.xlsx"  
overview\_data <- read\_excel(file\_path, sheet = "BUILDING\_OVERVIEW")  
  
file\_path <- "../DATA/BUILDING\_DISCIPLINE\_1819.xlsx"  
discipline\_data <- read\_excel(file\_path, sheet = "DISCIPLINE")

overview\_data <- dplyr::select(  
 overview\_data,  
 "Building IRN",  
 "Building Name",  
 "District Name",  
 "County",  
 "Region",  
 "Enrollment 2018-2019",  
 "Attendance Rate 2018-2019",  
 "Chronic Absenteeism Percent 2018-2019"  
)  
  
overview\_data[,6:8] <- sapply(overview\_data[,6:8],as.numeric)

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion  
  
## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

discipline\_data[ discipline\_data == "<10" ] <- "0"  
discipline\_data[,8:26] <- sapply(discipline\_data[,8:26],as.numeric)

## Warning in lapply(X = X, FUN = FUN, ...): NAs introduced by coercion

discipline\_data <- discipline\_data[,-26]  
discipline\_data <- subset(discipline\_data, discipline\_data$`Discpline Reason Description` == "Disobedient/Disruptive Behavior")

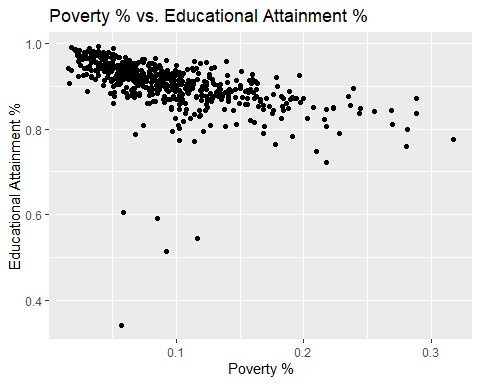
discipline\_data <- dplyr::select(  
 discipline\_data,  
 "Building IRN",  
 "Students Disciplined - Expulsions",  
 "Students Disciplined - Out-of-School Suspensions",  
 "Students Disciplined - In-School Suspensions"  
)

discipline\_data <- discipline\_data %>% mutate(  
 "total\_students\_discipline" = (rowSums(discipline\_data[,2:4])))  
  
joined\_df <- left\_join(overview\_data, discipline\_data, by = "Building IRN")

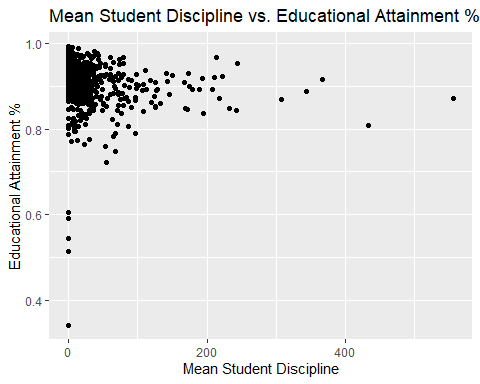
grouped\_df <- joined\_df %>%  
 group\_by(`District Name`, County, Region) %>%  
 summarise(mean\_enrollment = mean(`Enrollment 2018-2019`, na.rm = TRUE),  
 mean\_attendence = mean(`Attendance Rate 2018-2019`, na.rm = TRUE),  
 mean\_chronic\_absenteesim = mean(`Chronic Absenteeism Percent 2018-2019`, na.rm = TRUE),  
 mean\_expulsions = mean(`Students Disciplined - Expulsions`, na.rm = TRUE),  
 mean\_out\_of\_school\_suspensions = mean(`Students Disciplined - Out-of-School Suspensions`, na.rm = TRUE),  
 mean\_in\_school\_suspensions = mean(`Students Disciplined - In-School Suspensions`, na.rm = TRUE),  
 mean\_total\_students\_discipline = mean(`total\_students\_discipline`, na.rm = TRUE))

## `summarise()` has grouped output by 'District Name', 'County'. You can override using the `.groups` argument.

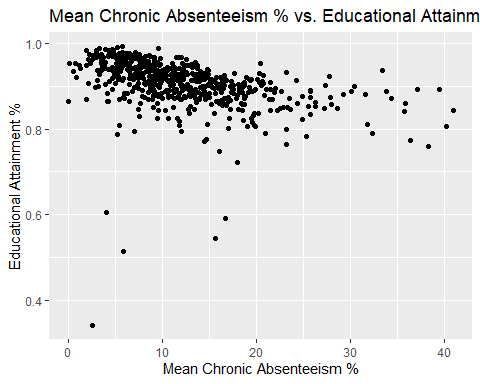
file\_path <- "../DATA/model\_data\_3\_30.csv"  
model\_data <- read.csv(file\_path)  
  
ggplot(model\_data, aes(x=poverty\_percentage, y=HS\_PLUS\_percentage)) + geom\_point() +   
 ggtitle("Poverty % vs. Educational Attainment %") +  
 xlab("Poverty %") +   
 ylab("Educational Attainment %")



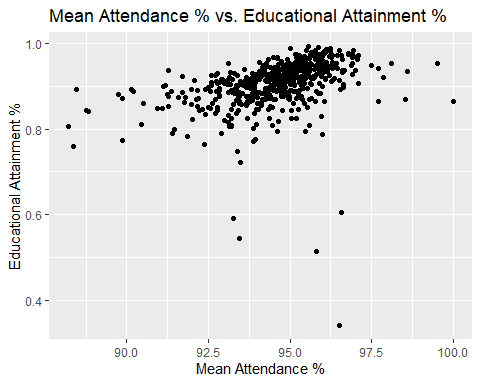
ggplot(model\_data, aes(x=mean\_total\_students\_discipline, y=HS\_PLUS\_percentage)) + geom\_point() +   
 ggtitle("Mean Student Discipline vs. Educational Attainment %") +  
 xlab("Mean Student Discipline") +   
 ylab("Educational Attainment %")



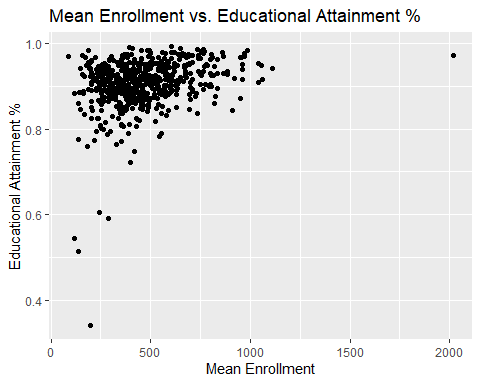
ggplot(model\_data, aes(x=mean\_chronic\_absenteeism, y=HS\_PLUS\_percentage)) + geom\_point() +   
 ggtitle("Mean Chronic Absenteeism % vs. Educational Attainment %") +  
 xlab("Mean Chronic Absenteeism %") +   
 ylab("Educational Attainment %")



ggplot(model\_data, aes(x=mean\_attendance, y=HS\_PLUS\_percentage)) + geom\_point() +   
 ggtitle("Mean Attendance % vs. Educational Attainment %") +  
 xlab("Mean Attendance %") +   
 ylab("Educational Attainment %")



ggplot(model\_data, aes(x=mean\_enrollment, y=HS\_PLUS\_percentage)) + geom\_point() +   
 ggtitle("Mean Enrollment vs. Educational Attainment %") +  
 xlab("Mean Enrollment") +   
 ylab("Educational Attainment %")



file\_path <- "../DATA/model\_data\_3\_30.csv"  
model\_data <- read.csv(file\_path)  
  
mod\_model\_data <- subset(model\_data, model\_data$HS\_PLUS\_percentage > 0.65)  
  
mod\_model\_data$HS\_PLUS\_percentage <- (((mod\_model\_data$HS\_PLUS\_percentage ^ 6) - 1) / 6)  
  
library(gam)

## Loading required package: splines

## Loading required package: foreach

##   
## Attaching package: 'foreach'

## The following objects are masked from 'package:purrr':  
##   
## accumulate, when

## Loaded gam 1.20

library(mgcv)

## Loading required package: nlme

##   
## Attaching package: 'nlme'

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

## This is mgcv 1.8-34. For overview type 'help("mgcv-package")'.

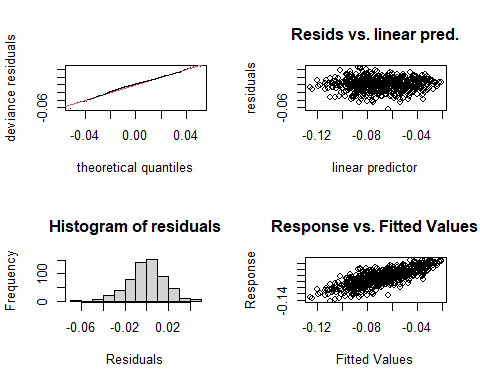
##   
## Attaching package: 'mgcv'

## The following objects are masked from 'package:gam':  
##   
## gam, gam.control, gam.fit, s

gam\_model\_1 <- gam(HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) + s(mean\_attendance) + s(mean\_enrollment) + s(mean\_total\_students\_discipline), data = mod\_model\_data)  
summary(gam\_model\_1)

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) +   
## s(mean\_attendance) + s(mean\_enrollment) + s(mean\_total\_students\_discipline)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0688202 0.0006799 -101.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(poverty\_percentage) 5.721 6.889 34.294 <2e-16 \*\*\*  
## s(mean\_chronic\_absenteeism) 2.773 3.463 2.880 0.0309 \*   
## s(mean\_attendance) 2.605 3.329 1.779 0.2807   
## s(mean\_enrollment) 4.177 5.096 10.318 <2e-16 \*\*\*  
## s(mean\_total\_students\_discipline) 1.636 2.042 0.482 0.6517   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.614 Deviance explained = 62.5%  
## GCV = 0.00028129 Scale est. = 0.00027275 n = 590

gam.check(gam\_model\_1)

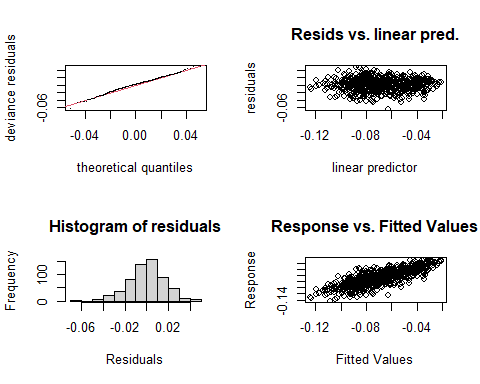


##   
## Method: GCV Optimizer: magic  
## Smoothing parameter selection converged after 4 iterations.  
## The RMS GCV score gradient at convergence was 1.215713e-07 .  
## The Hessian was not positive definite.  
## Model rank = 46 / 46   
##   
## Basis dimension (k) checking results. Low p-value (k-index<1) may  
## indicate that k is too low, especially if edf is close to k'.  
##   
## k' edf k-index p-value  
## s(poverty\_percentage) 9.00 5.72 1.01 0.54  
## s(mean\_chronic\_absenteeism) 9.00 2.77 0.98 0.29  
## s(mean\_attendance) 9.00 2.61 1.01 0.46  
## s(mean\_enrollment) 9.00 4.18 1.02 0.72  
## s(mean\_total\_students\_discipline) 9.00 1.64 1.01 0.65

gam\_model\_2 <- gam(HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) + s(mean\_attendance) + s(mean\_enrollment), data = mod\_model\_data)  
summary(gam\_model\_2)

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) +   
## s(mean\_attendance) + s(mean\_enrollment)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.06882 0.00068 -101.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(poverty\_percentage) 5.201 6.343 38.020 <2e-16 \*\*\*  
## s(mean\_chronic\_absenteeism) 2.734 3.415 2.966 0.0277 \*   
## s(mean\_attendance) 2.732 3.490 1.965 0.2467   
## s(mean\_enrollment) 4.242 5.169 11.210 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.613 Deviance explained = 62.3%  
## GCV = 0.00028041 Scale est. = 0.00027284 n = 590

gam.check(gam\_model\_2)

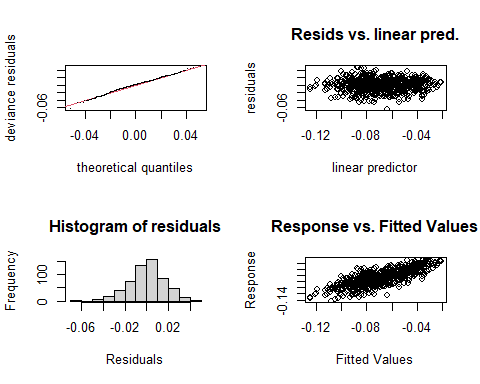


##   
## Method: GCV Optimizer: magic  
## Smoothing parameter selection converged after 4 iterations.  
## The RMS GCV score gradient at convergence was 3.410101e-09 .  
## The Hessian was positive definite.  
## Model rank = 37 / 37   
##   
## Basis dimension (k) checking results. Low p-value (k-index<1) may  
## indicate that k is too low, especially if edf is close to k'.  
##   
## k' edf k-index p-value  
## s(poverty\_percentage) 9.00 5.20 1.00 0.52  
## s(mean\_chronic\_absenteeism) 9.00 2.73 0.98 0.29  
## s(mean\_attendance) 9.00 2.73 1.01 0.56  
## s(mean\_enrollment) 9.00 4.24 1.02 0.74

gam\_model\_3 <- gam(HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) + s(mean\_enrollment) + s(mean\_total\_students\_discipline), data = mod\_model\_data)  
summary(gam\_model\_3)

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) +   
## s(mean\_enrollment) + s(mean\_total\_students\_discipline)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.0688202 0.0006834 -100.7 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(poverty\_percentage) 5.067 6.195 38.738 <2e-16 \*\*\*  
## s(mean\_chronic\_absenteeism) 2.381 3.026 7.144 0.0001 \*\*\*  
## s(mean\_enrollment) 4.143 5.058 10.786 <2e-16 \*\*\*  
## s(mean\_total\_students\_discipline) 2.153 2.710 0.977 0.4911   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.61 Deviance explained = 61.9%  
## GCV = 0.00028262 Scale est. = 0.00027556 n = 590

gam.check(gam\_model\_3)

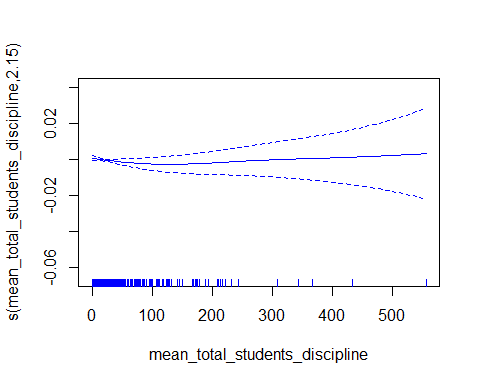
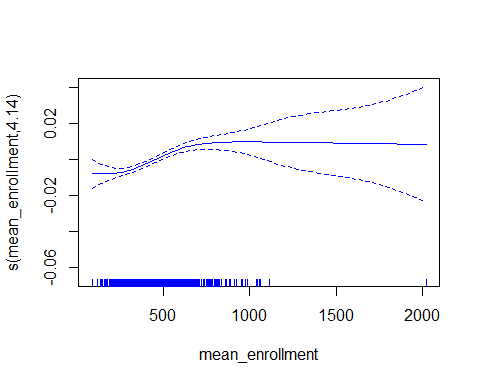
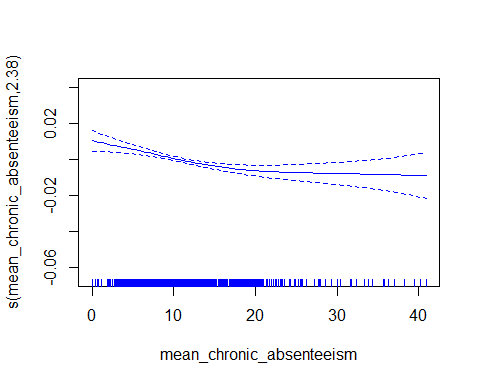
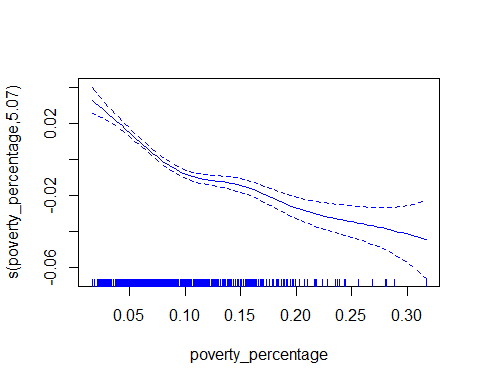


##   
## Method: GCV Optimizer: magic  
## Smoothing parameter selection converged after 4 iterations.  
## The RMS GCV score gradient at convergence was 9.066325e-08 .  
## The Hessian was positive definite.  
## Model rank = 37 / 37   
##   
## Basis dimension (k) checking results. Low p-value (k-index<1) may  
## indicate that k is too low, especially if edf is close to k'.  
##   
## k' edf k-index p-value  
## s(poverty\_percentage) 9.00 5.07 1.00 0.44  
## s(mean\_chronic\_absenteeism) 9.00 2.38 0.98 0.28  
## s(mean\_enrollment) 9.00 4.14 1.03 0.71  
## s(mean\_total\_students\_discipline) 9.00 2.15 1.01 0.52

anova(gam\_model\_1, gam\_model\_2, gam\_model\_3, test = "F")

## Analysis of Deviance Table  
##   
## Model 1: HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) +   
## s(mean\_attendance) + s(mean\_enrollment) + s(mean\_total\_students\_discipline)  
## Model 2: HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) +   
## s(mean\_attendance) + s(mean\_enrollment)  
## Model 3: HS\_PLUS\_percentage ~ s(poverty\_percentage) + s(mean\_chronic\_absenteeism) +   
## s(mean\_enrollment) + s(mean\_total\_students\_discipline)  
## Resid. Df Resid. Dev Df Deviance F Pr(>F)   
## 1 568.18 0.15604   
## 2 570.58 0.15664 -2.4015 -0.00060156 0.9184 0.41447   
## 3 572.01 0.15852 -1.4273 -0.00187929 4.8273 0.01682 \*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

plot(gam\_model\_3, se=TRUE, col="blue")



s