Satinitigan_Karl_HW4

Karl Satinitigan

2/16/2020

MACS30100

Non-linear regression

Egalitarianism and income

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------- tidyvers
```

```
## v ggplot2 3.2.1     v purrr   0.3.3
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   1.0.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------------ tidyverse_confl
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
library(tidymodels)
```

```
## Registered S3 method overwritten by 'xts':
##   method     from
##   as.zoo.xts zoo
```

```
## -- Attaching packages -------------------------------------------------------------------- tidymodel
```

```
## v broom     0.5.4     v recipes   0.1.9
## v dials     0.0.4     v rsample   0.0.5
## v infer     0.5.1     v yardstick 0.0.4
## v parsnip   0.0.5
```

```
## -- Conflicts ----------------------------------------------------------------------- tidymodels_confl
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x dials::margin()   masks ggplot2::margin()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
```

```r
library(rcfss)
library(knitr)
library(splines)
library(lattice)
library(here)
```

```
## here() starts at /Users/karl/Documents/UChicago/0 Computational Modeling/Problem Sets/Satinitigan_Ka
```

```r
library(patchwork)
library(margins)
library(ISLR)
library(boot)
```

```
##
## Attaching package: 'boot'

## The following object is masked from 'package:lattice':
##
##     melanoma
```

```r
library(readr)
library(glmnet)
```

```
## Loading required package: Matrix

##
## Attaching package: 'Matrix'

## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack

## Loaded glmnet 3.0-2
```

```r
library(caret)
```

```
##
## Attaching package: 'caret'

## The following objects are masked from 'package:yardstick':
##
##     precision, recall

## The following object is masked from 'package:purrr':
##
##     lift
```

```r
library(pls)
```

```
##
## Attaching package: 'pls'

## The following object is masked from 'package:caret':
##
##     R2

## The following object is masked from 'package:stats':
##
##     loadings
```

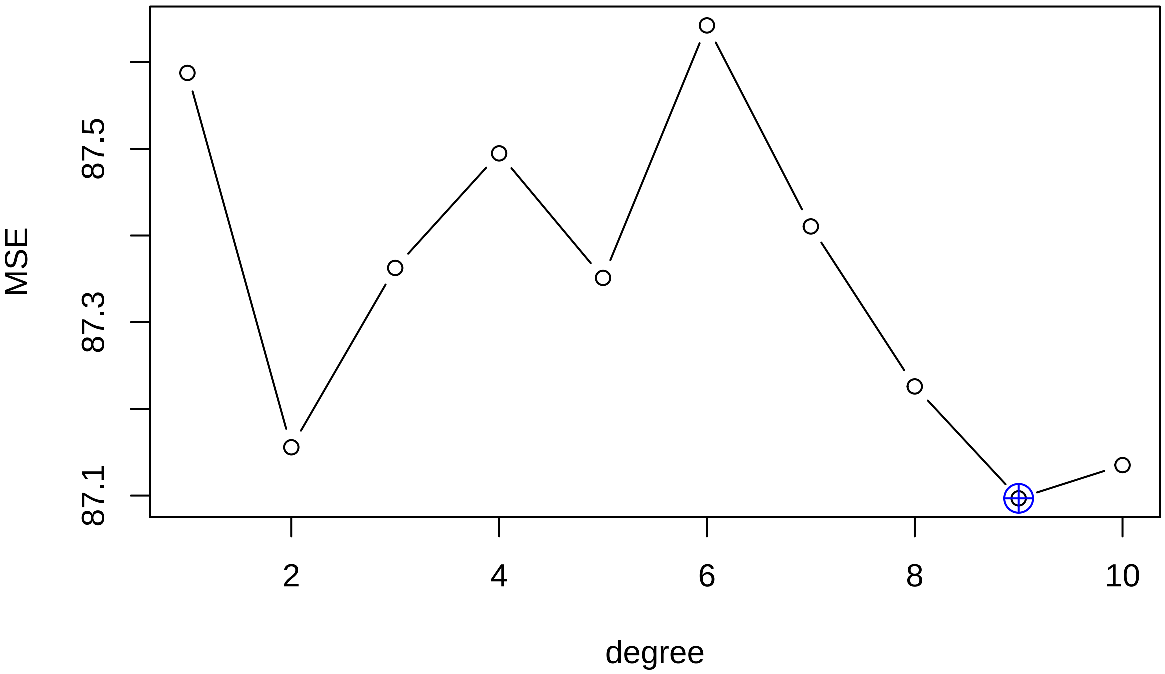
```r
set.seed(1234)
theme_set(theme_minimal())

gsstrain <- read_csv(url("https://raw.githubusercontent.com/ksatinitigan/problem-set-4/master/data/gss_
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   authoritarianism = col_double(),
##   childs = col_double(),
##   con_govt = col_double(),
##   egalit_scale = col_double(),
##   income06 = col_double(),
##   science_quiz = col_double(),
##   sibs = col_double(),
##   social_connect = col_double(),
##   tolerance = col_double(),
##   tvhours = col_double(),
##   wordsum = col_double()
## )

## See spec(...) for full column specifications.
```

```
gsstest <- read_csv(url("https://raw.githubusercontent.com/ksatinitigan/problem-set-4/master/data/gss_te
```

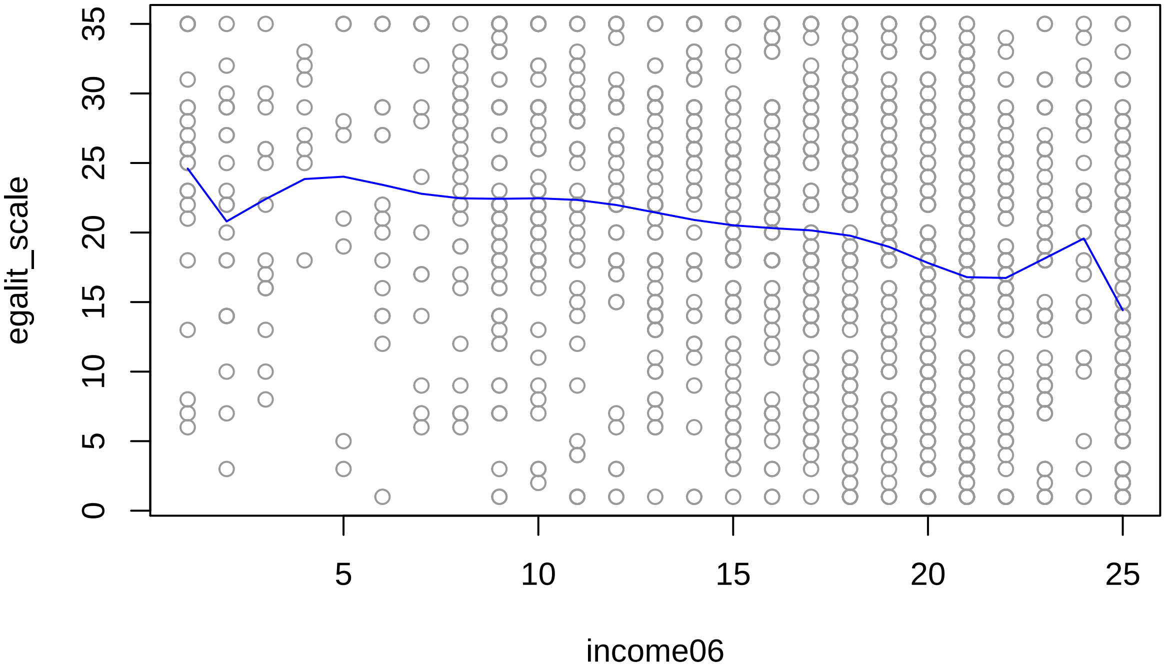
```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   authoritarianism = col_double(),
##   childs = col_double(),
##   con_govt = col_double(),
##   egalit_scale = col_double(),
##   income06 = col_double(),
##   science_quiz = col_double(),
##   sibs = col_double(),
##   social_connect = col_double(),
##   tolerance = col_double(),
##   tvhours = col_double(),
##   wordsum = col_double()
## )
## See spec(...) for full column specifications.
```

```
### Polynomial regression

cvMSE <- NA
for (i in 1:10){
  glmfit <- glm(egalit_scale ~ poly(income06, i), data = gsstrain)
  cvMSE[i] <- cv.glm(gsstrain, glmfit, K = 10)$delta[1]
}

plot(1:10, cvMSE, xlab = "degree", ylab = "MSE", type = "b")
mindegree <- which.min(cvMSE)
points(mindegree, cvMSE[mindegree], col = "blue", cex = 2, pch = 10)
```

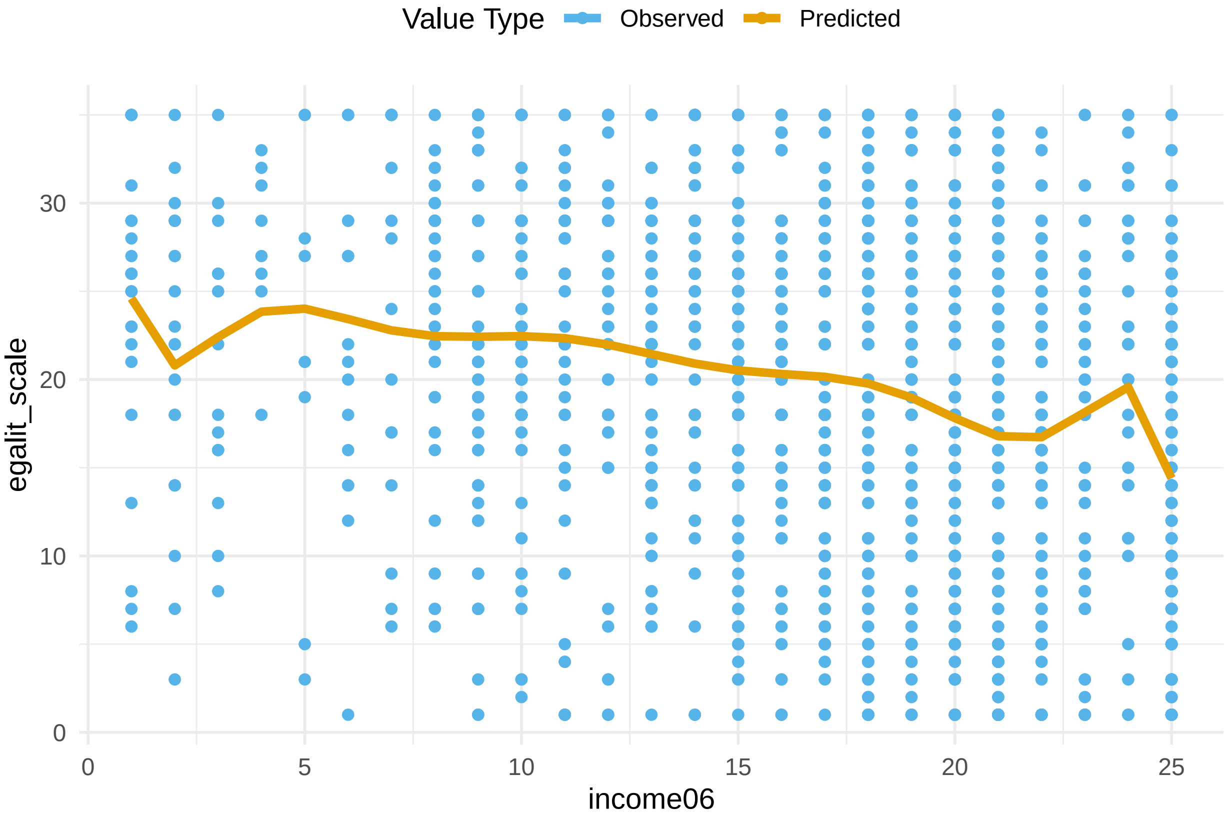


d = 9

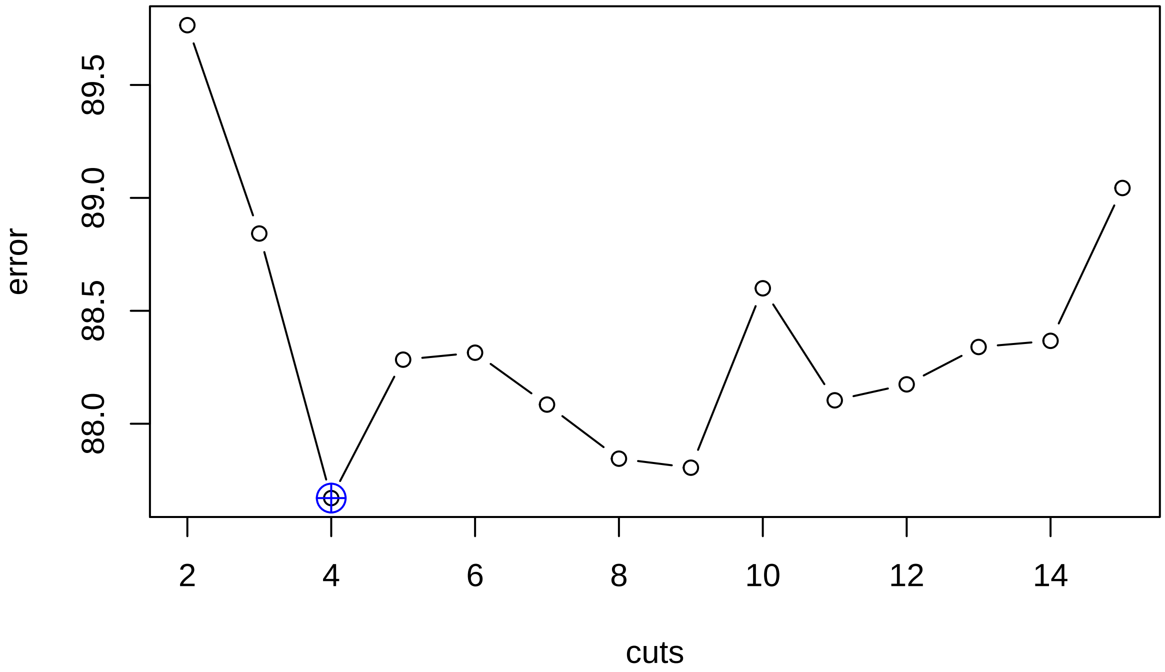
```
plot(egalit_scale ~ income06, data = gsstrain, col = "grey60")
incomelim <- range(gsstrain$income06)
incomegrid <- seq(from = incomelim[1], to = incomelim[2])
polyfit <- lm(egalit_scale ~ poly(income06, 9), data = gsstrain)
polypred <- predict(polyfit, newdata = list(income06 = incomegrid))
lines(incomegrid, polypred, col="blue")
```



```
gsstrain %>%
    mutate(pred = predict(polyfit, gsstrain)) %>%
    ggplot() +
    geom_point(aes(income06, egalit_scale, col = 'blue')) +
    geom_line(aes(income06, pred, col = 'goldenrod2'), size = 1.5) +
    scale_color_manual(name = 'Value Type',
                       labels = c('Observed', 'Predicted'),
                       values = c('#56B4E9', '#E69F00')) +
    labs(x = 'income06', y = 'egalit_scale') +
    theme(legend.position = 'top')
```

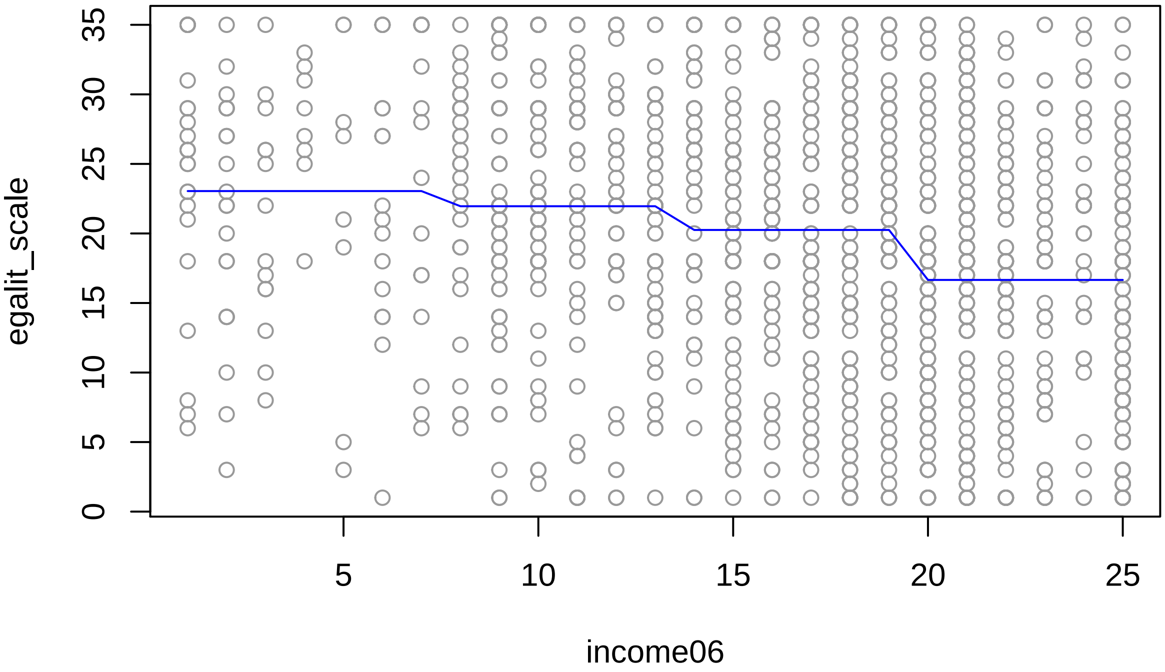


```
### Step function

cvError <- NA
for (i in 2:15) {
  gsstrain$income06cut <- cut(gsstrain$income06, i)
  lmfit <- glm(egalit_scale ~ income06cut, data = gsstrain)
  cvError[i] <- cv.glm(gsstrain, lmfit, K = 10)$delta[1]
}
plot(2:15, cvError[-1], xlab = "cuts", ylab = "error", type = "b")
points(x = which.min(cvError), y =  min(cvError, na.rm = TRUE), col = "blue", cex = 2, pch = 10)
```

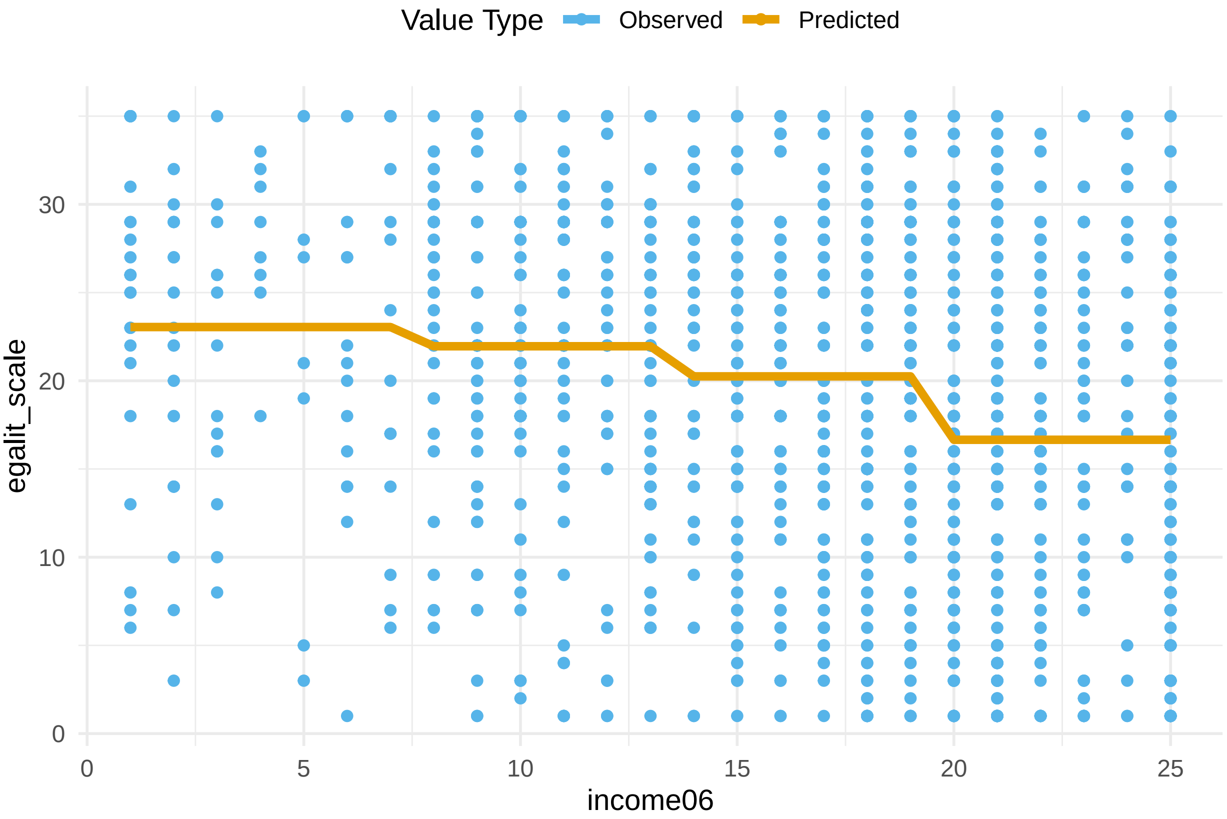


cuts = 4

```r
plot(egalit_scale ~ income06, data = gsstrain, col = "grey60")
stepfit <- glm(egalit_scale ~ cut(income06, 4), data = gsstrain)
steppred <- predict(stepfit, list(income06 = incomegrid))
lines(incomegrid, steppred, col="blue")
```



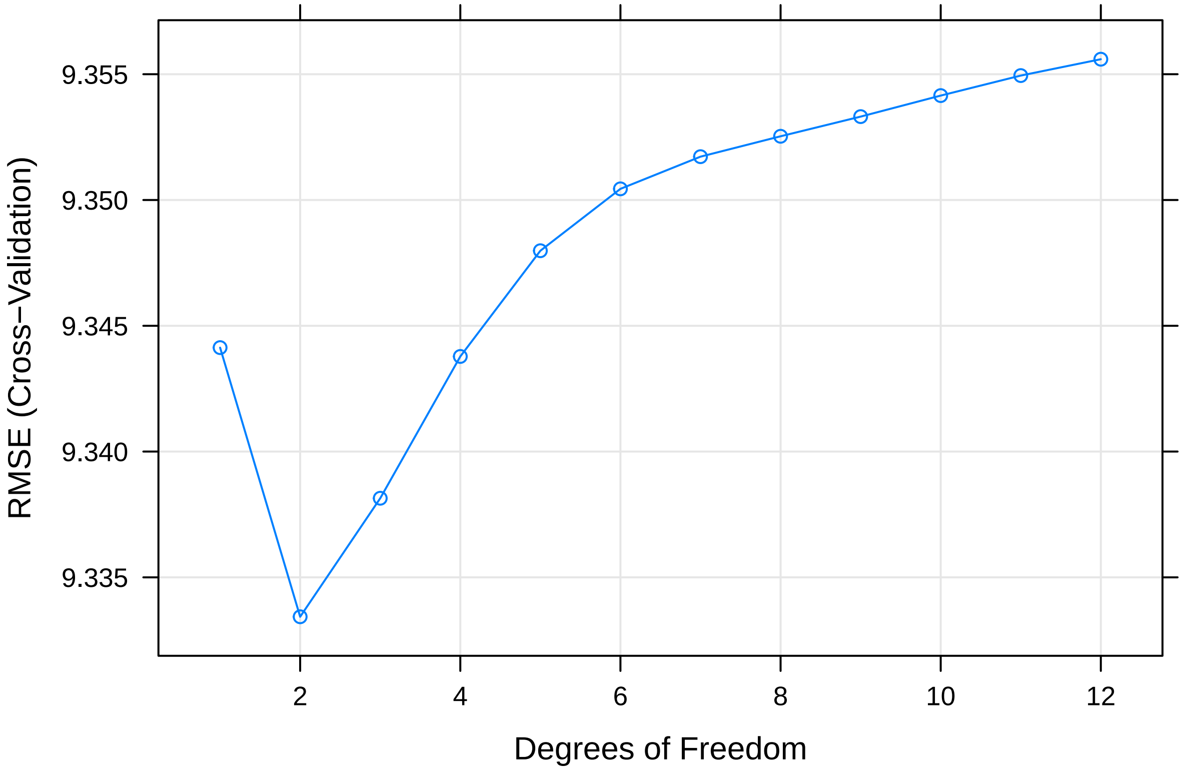
```r
gsstrain %>%
    mutate(pred = predict(stepfit, gsstrain)) %>%
    ggplot() +
    geom_point(aes(income06, egalit_scale, col = 'blue')) +
    geom_line(aes(income06, pred, col = 'goldenrod2'), size = 1.5) +
    scale_color_manual(name = 'Value Type',
                       labels = c('Observed', 'Predicted'),
                       values = c('#56B4E9', '#E69F00')) +
    labs(x = 'income06', y = 'egalit_scale') +
    theme(legend.position = 'top')
```



```
### Natural regression spline

spline <- train(egalit_scale ~ income06, data = gsstrain,
                method = "gamSpline",
                trControl = trainControl(method = "cv", number = 10),
                tuneGrid = expand.grid(df = seq(1, 12, 1)))
```

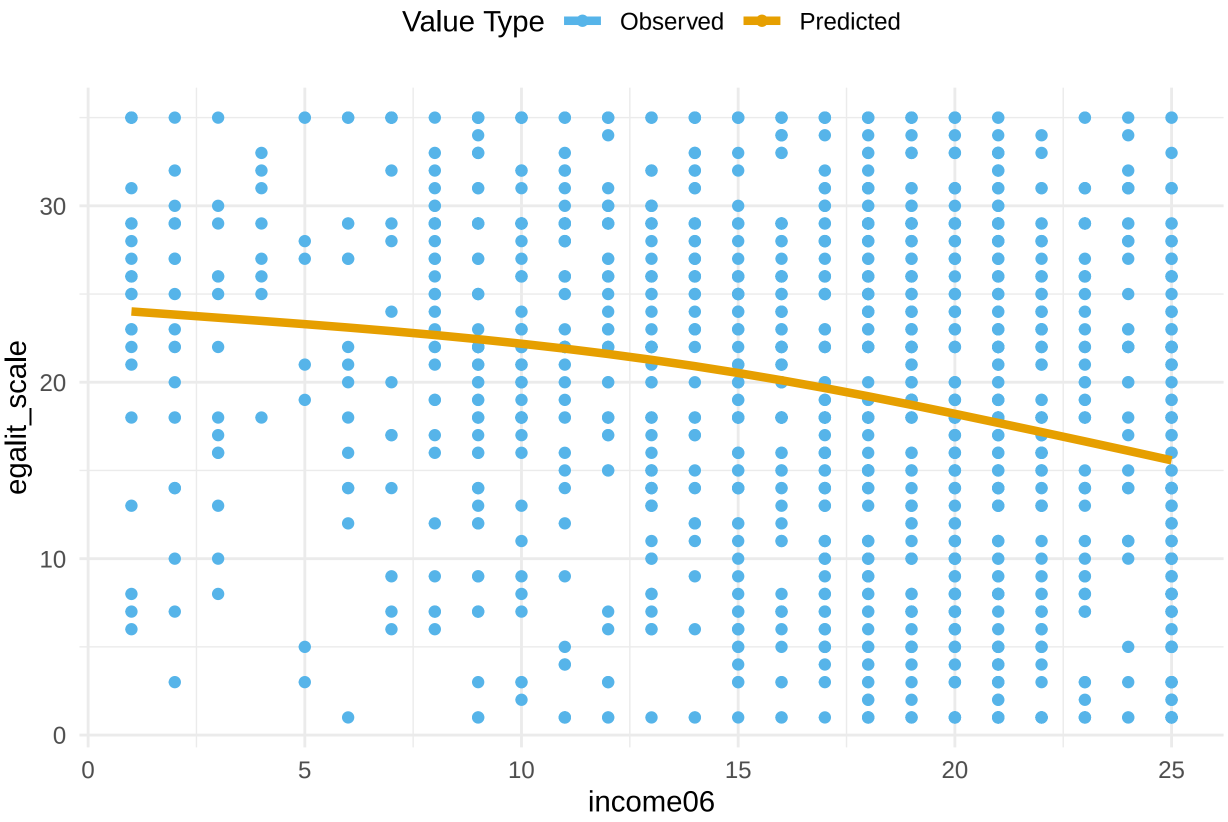
```
## Loading required package: gam

## Loading required package: foreach

##
## Attaching package: 'foreach'

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

## Loaded gam 1.16.1
```

```
income_lm <- train(egalit_scale ~ income06, data = gsstrain,
                method = "lm",
                trControl = trainControl(method = "cv", number = 10))
plot(spline)
```



d = 2

```
gsstrain %>%
    mutate(pred = predict(spline, gsstrain)) %>%
    ggplot() +
    geom_point(aes(income06, egalit_scale, col = 'blue')) +
    geom_line(aes(income06, pred, col = 'goldenrod2'), size = 1.5) +
    scale_color_manual(name = 'Value Type',
                       labels = c('Observed', 'Predicted'),
                       values = c('#56B4E9', '#E69F00')) +
    labs(x = 'income06', y = 'egalit_scale') +
    theme(legend.position = 'top')
```



```
## Egalitarianism and everything


gsstrainEv <- read_csv(url("https://raw.githubusercontent.com/ksatinitigan/problem-set-4/master/data/gss
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   authoritarianism = col_double(),
##   childs = col_double(),
##   con_govt = col_double(),
##   egalit_scale = col_double(),
##   income06 = col_double(),
##   science_quiz = col_double(),
##   sibs = col_double(),
##   social_connect = col_double(),
##   tolerance = col_double(),
##   tvhours = col_double(),
##   wordsum = col_double()
## )
```

```
## See spec(...) for full column specifications.
```

```
gsstestEv <- read_csv(url("https://raw.githubusercontent.com/ksatinitigan/problem-set-4/master/data/gss_
```

```
## Parsed with column specification:
## cols(
##   .default = col_character(),
##   age = col_double(),
##   authoritarianism = col_double(),
##   childs = col_double(),
##   con_govt = col_double(),
##   egalit_scale = col_double(),
##   income06 = col_double(),
##   science_quiz = col_double(),
##   sibs = col_double(),
##   social_connect = col_double(),
##   tolerance = col_double(),
##   tvhours = col_double(),
##   wordsum = col_double()
## )
## See spec(...) for full column specifications.
```

```r
gsstrainEvmain <- gsstrainEv$egalit_scale
gsstrainEvmod <- dummyVars(egalit_scale ~., gsstrainEv)
gsstrainEvmat <- predict(gsstrainEvmod, newdata = gsstrainEv)
gsstrainEv <- data.frame(gsstrainEvmat)
gsstrainEvmain <- cbind(gsstrainEvmain, gsstrainEv)
gsstrainEvmain <- rename(gsstrainEvmain, egalit_scale = gsstrainEvmain)
parameters1 <- preProcess(gsstrainEvmain, method = c("center", "scale", "zv"))
gsstrainEvmain <- predict(parameters1, gsstrainEvmain)
gsstrainEvmain <- subset(gsstrainEvmain, select = -c(religORTHODOX.CHRISTIAN))

gsstestEvmain <- gsstestEv$egalit_scale
gsstestEvmod <- dummyVars(egalit_scale ~., gsstestEv)
gsstestEvmat <- predict(gsstestEvmod, newdata = gsstestEv)
gsstestEv <- data.frame(gsstestEvmat)
gsstestEvmain <- cbind(gsstestEvmain, gsstestEv)
gsstestEvmain <- rename(gsstestEvmain, egalit_scale = gsstestEvmain)
parameters2 <- preProcess(gsstestEvmain, method = c("center", "scale", "zv"))
gsstestEvmain <- predict(parameters2, gsstestEvmain)


### Linear regression

linear <- lm(egalit_scale ~., gsstrainEvmain)
linearpred <- predict(linear, gsstestEvmain)
linearMSE <- mean((gsstestEvmain$egalit_scale - linearpred)^2)
linearMSE
```

```
## [1] 0.6933239
```

Test MSE is 0.6933239

```r
### Elastic net regression

gsstrainx <- model.matrix(egalit_scale ~ ., gsstrainEvmain)[, -1]
gsstrainy <- gsstrainEvmain$egalit_scale

gsstestx <- model.matrix(egalit_scale ~., gsstestEvmain)[, -1]
gsstesty <- gsstestEvmain$egalit_scale

for (i in seq(0, 1, .1))
elasticCV <- cv.glmnet(gsstrainx, gsstrainy, alpha=i)
```

```r
bestlamelastic = elasticCV$lambda.min

elastic <- glmnet(gsstrainx, gsstrainy, alpha=1, lambda=bestlamelastic)
elastic$beta
```

```
## 134 x 1 sparse Matrix of class "dgCMatrix"
##                                         s0
## age                          -7.493173e-02
## attend.Once.yr                .
## attend.Once.wk                .
## attend2.3.times..mo           .
## attendEvery.wk               -1.397717e-03
## attendNever                   .
## attendNrly.evry.wk            4.960351e-03
## attendOnce.mo                 .
## attendOnce.yr                 .
## attendSev.times.yr            .
## authoritarianism              .
## blackNo                      -2.321410e-02
## blackYes                      1.467229e-14
## bornNO                        .
## bornYES                       .
## childs                        2.538804e-02
## colathALLOWED                 .
## colathNOT.ALLOWED             .
## colracALLOWED                 .
## colracNOT.ALLOWED             .
## colcomFIRED                   .
## colcomNOT.FIRED               .
## colmilALLOWED                 .
## colmilNOT.ALLOWED             .
## colhomoALLOWED               -5.018626e-03
## colhomoNOT.ALLOWED            .
## colmslmNot.allowed            .
## colmslmYes..allowed           .
## con_govt                      .
## degree.HS                     8.045364e-03
## degreeBachelor.deg           -7.019781e-02
## degreeGraduate.deg            .
## degreeHS                      .
## degreeJunior.Coll            -1.184257e-02
## evangelicalHigh               .
## evangelicalLow                .
## evangelicalMod                .
## grassLEGAL                    7.263818e-02
## grassNOT.LEGAL               -2.715536e-14
## happyNOT.TOO.HAPPY            2.175888e-03
## happyPRETTY.HAPPY             .
## happyVERY.HAPPY              -3.453241e-03
## hispanic_2No                  .
## hispanic_2Yes                 .
## homosexALMST.ALWAYS.WRG       .
## homosexALWAYS.WRONG           .
## homosexNOT.WRONG.AT.ALL       .
## homosexSOMETIMES.WRONG        .
## income06                     -7.886866e-02
## maritalDivorced               .
## maritalMarried                .
## maritalNever.married          .
## maritalSeparated              .
## maritalWidowed               -4.634981e-04
## modeIN.PERSON                 .
## modeOVER.THE.PHONE            .
## newsEVERYDAY                  .
## newsFEW.TIMES.A.WEEK          .
## newsLESS.THAN.ONCE.WK         .
## newsNEVER                     5.573449e-03
## newsONCE.A.WEEK               .
## owngunNO                      4.434837e-02
## owngunREFUSED                 .
## owngunYES                     .
## partyid_3Dem                  8.347713e-02
## partyid_3Ind                  .
## partyid_3Rep                 -7.206378e-02
## polviewsConserv              -7.537283e-02
## polviewsExtrmCons            -6.393625e-02
## polviewsExtrmLib              5.906472e-02
## polviewsLiberal               1.262032e-01
## polviewsModerate              .
## polviewsSlghtCons            -1.692359e-02
## polviewsSlghtLib              6.651629e-02
## pornlaw2Illegal.to.all        .
## pornlaw2Not.illegal.to.all    .
## prayLT.ONCE.A.WEEK            .
## prayNEVER                     .
## prayONCE.A.DAY                .
## prayONCE.A.WEEK               .
## praySEVERAL.TIMES.A.DAY       .
## praySEVERAL.TIMES.A.WEEK      .
## pres08McCain                 -2.033679e-01
## pres08Obama                   7.875645e-15
## reborn_rNo                    .
## reborn_rYes                   .
## religBUDDHISM                 .
## religCATHOLIC                 .
## religCHRISTIAN                .
## religHINDUISM                -8.173875e-03
## religINTER.NONDENOMINATIONAL  .
## religJEWISH                   .
## religMOSLEM.ISLAM             .
## religNATIVE.AMERICAN          .
## religNONE                     .
## religOTHER                    .
## religOTHER.EASTERN            .
## religPROTESTANT               .
## science_quiz                 -1.461911e-02
## sexFemale                     4.395364e-02
## sexMale                       .
## sibs                          2.592854e-02
## social_connect                .
## social_cons3Conserv           .
## social_cons3Liberal           .
## social_cons3Mod               .
## southNonsouth                 .
## southSouth                    .
## spend3Conserv                -4.426462e-02
## spend3Liberal                 3.506361e-02
## spend3Mod                     .
## teensexALMST.ALWAYS.WRG       .
## teensexALWAYS.WRONG           .
## teensexNOT.WRONG.AT.ALL       .
## teensexSOMETIMES.WRONG        .
## tolerance                    -6.007539e-02
## tvhours                       3.908390e-02
## vetyears2.TO.4.YEARS          .
## vetyearsLESS.THAN.2.YRS       .
## vetyearsMORE.THAN.4.YRS       .
## vetyearsNONE                  5.895945e-03
## wordsum                       .
## zodiacAQUARIUS                1.709432e-03
## zodiacARIES                   .
## zodiacCANCER                  .
## zodiacCAPRICORN               .
## zodiacGEMINI                  .
## zodiacLEO                     .
## zodiacLIBRA                   .
## zodiacPISCES                  .
## zodiacSAGITTARIUS             .
## zodiacSCORPIO                -2.312972e-03
## zodiacTAURUS                  .
## zodiacVIRGO                   5.821476e-03
```

```r
predictelastic <- predict(elastic, s=bestlamelastic, newx = gsstestx)
elasticMSE <- mean((predictelastic - gsstesty)^2)
elasticMSE
```

```
## [1] 0.6716835
```

Test MSE is 0.6716835

```r
### Principal component regression

PCRfit <- train(egalit_scale ~., data = gsstrainEvmain,
                method="pcr",
                scale = TRUE,
                trControl = trainControl("cv", number = 10),
                tuneLength = 10
                )
PCRfit
```

```
## Principal Component Analysis
##
## 1481 samples
##  134 predictor
```

```
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1334, 1332, 1333, 1333, 1333, 1333, ...
## Resampling results across tuning parameters:
##
##   ncomp  RMSE       Rsquared    MAE
##    1     0.9899626  0.01969522  0.8229787
##    2     0.8577424  0.26541727  0.6955589
##    3     0.8338368  0.30593300  0.6666252
##    4     0.8334023  0.30661140  0.6662629
##    5     0.8333524  0.30681542  0.6660915
##    6     0.8283919  0.31510024  0.6622930
##    7     0.8271728  0.31663633  0.6609357
##    8     0.8278173  0.31557526  0.6614885
##    9     0.8269553  0.31652529  0.6604562
##   10     0.8276962  0.31518942  0.6610947
##
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was ncomp = 9.
```

```
predictPCR <- predict(PCRfit, gsstestEvmain)
```

Components = 9

```
### Partial least squares regression

PLSfit <- train(egalit_scale ~., data = gsstrainEvmain,
                method = "pls",
                scale = TRUE,
                trControl = trainControl("cv", number = 10),
                tuneLength = 10)
PLSfit
```

```
## Partial Least Squares
##
## 1481 samples
##  134 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1332, 1333, 1333, 1334, 1333, 1332, ...
## Resampling results across tuning parameters:
##
##   ncomp  RMSE       Rsquared   MAE
##    1     0.8434902  0.2896075  0.6685268
##    2     0.8222004  0.3260621  0.6489407
##    3     0.8212299  0.3310959  0.6468318
##    4     0.8252532  0.3269075  0.6488906
##    5     0.8292209  0.3221465  0.6510618
##    6     0.8287369  0.3231171  0.6507819
##    7     0.8275695  0.3248878  0.6499408
##    8     0.8270338  0.3256262  0.6498263
##    9     0.8269436  0.3257067  0.6498385
##   10     0.8270934  0.3255213  0.6498679
##
```

```
## RMSE was used to select the optimal model using the smallest value.
## The final value used for the model was ncomp = 3.
```

```
predictPLS <- predict(PLSfit, gsstestEvmain)
```

Components = 2