StatMod3

Igor

24 мая 2019 г

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.5.3

library(memisc)

## Warning: package 'memisc' was built under R version 3.5.3

## Loading required package: lattice

## Loading required package: MASS

## Warning: package 'MASS' was built under R version 3.5.3

##   
## Attaching package: 'memisc'

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

## The following objects are masked from 'package:stats':  
##   
## contr.sum, contr.treatment, contrasts

## The following object is masked from 'package:base':  
##   
## as.array

library(DescTools)

## Warning: package 'DescTools' was built under R version 3.5.3

##   
## Attaching package: 'DescTools'

## The following object is masked from 'package:memisc':  
##   
## %nin%

library(broom)

## Warning: package 'broom' was built under R version 3.5.3

library(caTools)

## Warning: package 'caTools' was built under R version 3.5.3

library(lmtest)

## Warning: package 'lmtest' was built under R version 3.5.3

## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.5.3

##   
## Attaching package: 'zoo'

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

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.5.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:memisc':  
##   
## collect, recode, rename, syms

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

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

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

library(readxl)

## Warning: package 'readxl' was built under R version 3.5.3

housing <- read.csv("C:/Users/Igor/Downloads/housing.csv")  
housing <- housing[1:150,]  
set.seed(56)  
split <- sample.split(housing$median\_house\_value,SplitRatio = 0.75)  
train <- subset(housing,split == TRUE)  
test <- subset(housing,split == FALSE)  
model\_1 <- lm(data=train, median\_house\_value ~ median\_income)  
summary(model\_1)

##   
## Call:  
## lm(formula = median\_house\_value ~ median\_income, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -231671 -35386 -1114 35023 373867   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 80085 10928 7.329 4.14e-11 \*\*\*  
## median\_income 37034 2792 13.264 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 66250 on 110 degrees of freedom  
## Multiple R-squared: 0.6153, Adjusted R-squared: 0.6118   
## F-statistic: 175.9 on 1 and 110 DF, p-value: < 2.2e-16

summary(lm(data=train,median\_house\_value ~ median\_income))$r.squared

## [1] 0.6152979

summary(lm(data=train,median\_house\_value ~ median\_income))$adj.r.squared

## [1] 0.6118006

summary(lm(data=train,median\_house\_value ~ median\_income))$sigma^2

## [1] 4389014560

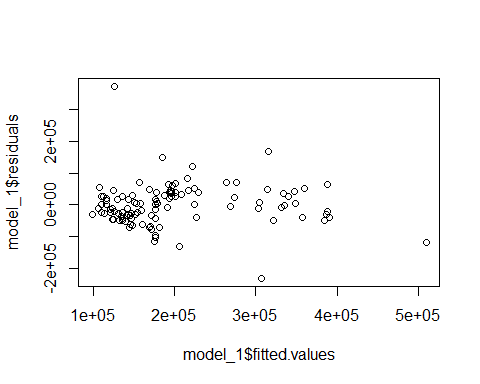
glance(model\_1)

## # A tibble: 1 x 11  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <dbl>  
## 1 0.615 0.612 66250. 176. 1.46e-24 2 -1401. 2808.  
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>

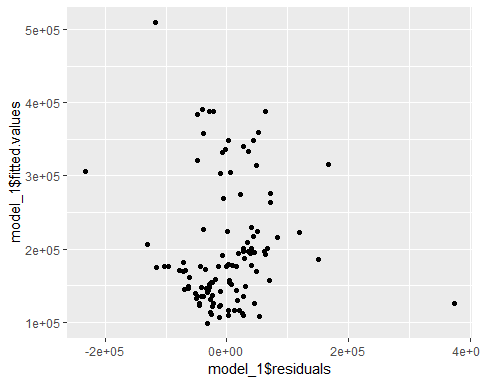
**Вывод об оценке качества подгонки**: Коэффициент детерминации равен 0.6153, скорректированный равен 0.6118, что является средним показателем. Остаточная дисперсия имеет довольно большое значение. Можно сделать вывод о том, что построенная нами регрессионная модель является недостаточно точной.

**Проверка различных параметров относительно параметров уравнения:** F-критерий: так как p-value<0.05, то гипотезу об отсутствии линейной связи отвергаем. t-критерий: так как p-value<0.05, то гипотезу о незначимости фактора X отвергаем.

plot(model\_1$fitted.values,model\_1$residuals)



qplot(y=model\_1$fitted.values,x=model\_1$residuals)



**Вывод о случайном характере остатков**: судя по графику остатки носят случайный характер, так как на графике нет направленности в расположении точек E.

mean(model\_1$residuals)

## [1] -7.522459e-12

a<-mean(model\_1$residuals)  
b<-sd(model\_1$residuals,na.rm = FALSE)  
n<-sqrt(110)  
tm<-a/b\*n  
str(tm)

## num -1.2e-15

qt(0.95,42125)

## [1] 1.64489

**Вывод о равенстве матожидания остатков равна нулю:** так как |tрасч| < tтабл, то принимается нулевая гипотеза о равенстве матожидания остатков нулю.

bptest(model\_1)

##   
## studentized Breusch-Pagan test  
##   
## data: model\_1  
## BP = 0.051272, df = 1, p-value = 0.8209

**Вывод**: так как p-value > 0.05, то гипотеза о гомоскедастичности остатков принимается.

cor.test(train$median\_income,model\_1$residuals,method="spearman")

## Warning in cor.test.default(train$median\_income, model\_1$residuals, method  
## = "spearman"): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: train$median\_income and model\_1$residuals  
## S = 184760, p-value = 0.02561  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.2109035

**Вывод**: так как p-value < 0.05, то гипотеза о гомоскедастичности остатков отвергается.

bgtest(model\_1)

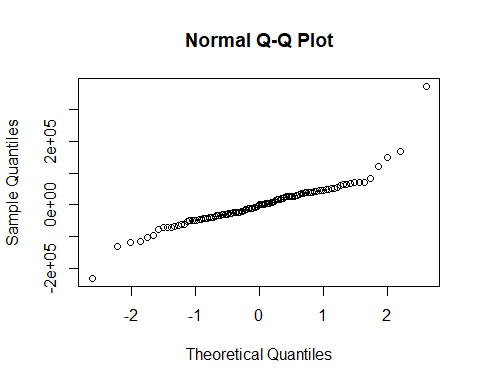
##   
## Breusch-Godfrey test for serial correlation of order up to 1  
##   
## data: model\_1  
## LM test = 18.912, df = 1, p-value = 1.369e-05

dwtest(model\_1)

##   
## Durbin-Watson test  
##   
## data: model\_1  
## DW = 1.1828, p-value = 4.343e-06  
## alternative hypothesis: true autocorrelation is greater than 0

**Вывод**: поскольку значение p-value < 0.05 нулевая гипотеза об отсутствии автокорреляции остатков отвергается.

qqnorm(model\_1$residuals)



library(sm)

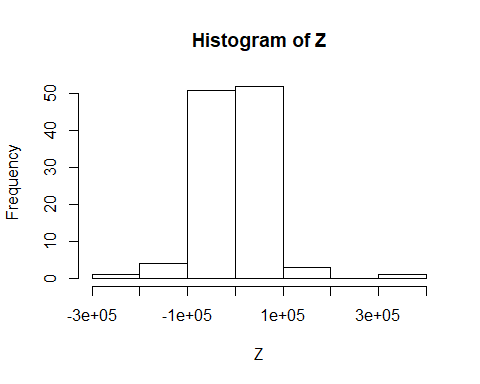
## Warning: package 'sm' was built under R version 3.5.3

## Package 'sm', version 2.2-5.6: type help(sm) for summary information

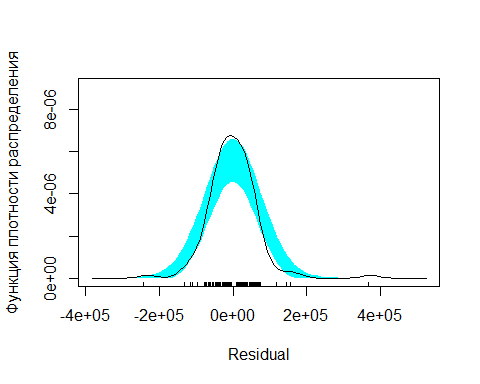
##   
## Attaching package: 'sm'

## The following object is masked from 'package:MASS':  
##   
## muscle

Z<-model\_1$residuals  
hist(Z)



sm.density(model\_1$residuals,model="Normal", xlab="Residual",ylab="Функция плотности распределения")



library(nortest)

## Warning: package 'nortest' was built under R version 3.5.2

lillie.test(model\_1$residuals)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: model\_1$residuals  
## D = 0.10072, p-value = 0.007121

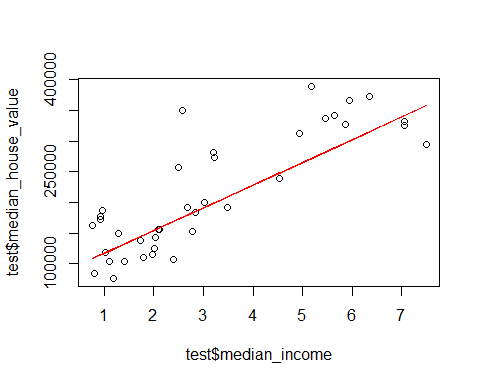
**Вывод**: так как значение p-value меньше 0.05, нулевая гипотеза о согласии распределения остатков с нормальным законом распределения отвергается.

sh<-model\_1$residuals[1:4990]  
View(sh)  
shapiro.test(sh)

##   
## Shapiro-Wilk normality test  
##   
## data: sh  
## W = 0.87827, p-value = 4.098e-08

**Вывод**: так как значение p-value меньше 0.05, нулевая гипотеза о согласии распределения остатков с нормальным законом распределения отвергается.

test$model\_1<-predict(model\_1,test)  
#View(test$model\_1)  
plot(test$median\_income,test$median\_house\_value)  
lines(test$median\_income,predict(model\_1,test),col='red')



**Выводы о корректировке модели**:

1. В нашей модели можно наблюдать автокорреляцию в остатках. То есть имеется зависимость случайных компонент наблюдений с различными номерами. Это приводит к получению неэффективных оценок.

2. Наши остатки не распределенные нормально и это приводит к получению несостоятельных оценок, и как следствие приводящих к ненадежным прогнозам.

insurance <- read.csv("C:/Users/Igor/Downloads/LifeExpectancyData2.csv")  
#View(housing)  
names(insurance)

## [1] "Country" "Year"   
## [3] "Status" "Life.expectancy"   
## [5] "Adult.Mortality" "infant.deaths"   
## [7] "Alcohol" "percentage.expenditure"   
## [9] "Hepatitis.B" "Measles"   
## [11] "BMI" "under.five.deaths"   
## [13] "Polio" "Total.expenditure"   
## [15] "Diphtheria" "HIV.AIDS"   
## [17] "GDP" "Population"   
## [19] "thinness..1.19.years" "thinness.5.9.years"   
## [21] "Income.composition.of.resources" "Schooling"

insurance2<- select(insurance,-"Country",-"Year",-"Status",-"thinness..1.19.years" ,-"thinness.5.9.years",-"percentage.expenditure",-"Total.expenditure",-"under.five.deaths",-"Alcohol",-"Hepatitis.B",-"Measles",-"Polio",-"Diphtheria",-"Population",-"BMI", -"GDP",-"infant.deaths")

names(insurance2)

## [1] "Life.expectancy" "Adult.Mortality"   
## [3] "HIV.AIDS" "Income.composition.of.resources"  
## [5] "Schooling"

insurance2 <- insurance2[1:500,]

set.seed(1821)  
split<-sample.split(insurance2$Life.expectancy,SplitRatio = 0.75)  
train<-subset(insurance2,split==TRUE)  
test<-subset(insurance2,split==FALSE)  
train\_y<-train$Life.expectancy  
train\_x<-train%>%select(-"Life.expectancy")%>%data.matrix()  
test\_x<-test%>%select(-"Life.expectancy")%>%data.matrix()

m\_lm <-lm(Life.expectancy~.,data=train)  
summary(m\_lm)

##   
## Call:  
## lm(formula = Life.expectancy ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.9782 -2.2779 0.3549 2.1270 12.1295   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 56.583676 1.081424 52.323 < 2e-16 \*\*\*  
## Adult.Mortality -0.028970 0.002348 -12.337 < 2e-16 \*\*\*  
## HIV.AIDS -0.349026 0.060296 -5.789 1.40e-08 \*\*\*  
## Income.composition.of.resources 8.653785 1.597218 5.418 1.02e-07 \*\*\*  
## Schooling 1.077960 0.099398 10.845 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.79 on 418 degrees of freedom  
## Multiple R-squared: 0.834, Adjusted R-squared: 0.8324   
## F-statistic: 525.1 on 4 and 418 DF, p-value: < 2.2e-16

test$predict<-predict(m\_lm,test)

View(test)

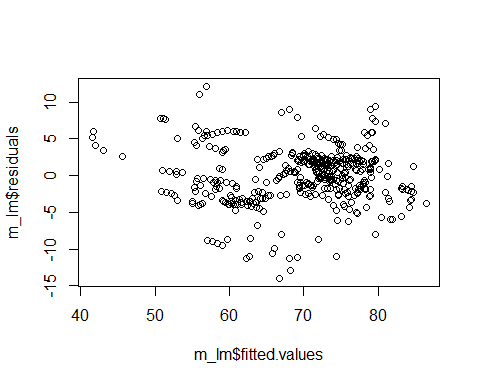
glance(m\_lm)

## # A tibble: 1 x 11  
## r.squared adj.r.squared sigma statistic p.value df logLik AIC  
## <dbl> <dbl> <dbl> <dbl> <dbl> <int> <dbl> <dbl>  
## 1 0.834 0.832 3.79 525. 1.71e-161 5 -1161. 2335.  
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>

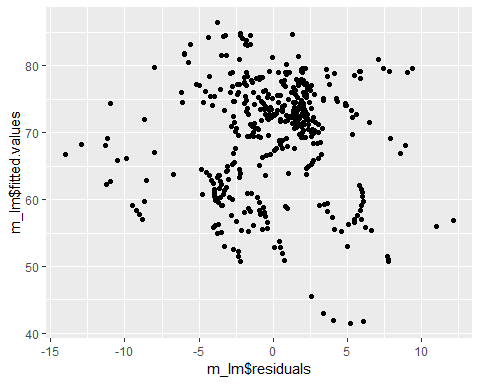
**Вывод об оценке качества подгонки:** Коэффициент скорректированной детерминации равен 0.832, что является хорошим показателем. Остаточная дисперсия имеет небольшое значение и это тоже хорошо. Можно сделать вывод о том, что построенная нами регрессионная модель является достаточно точной.

**Проверка различных параметров относительно параметров уравнения:** F-критерий: так как p-value<0.05, то гипотезу об отсутствии линейной связи отвергаем. t-критерий: так как p-value<0.05, то гипотезу о незначимости фактора X отвергаем.

plot(m\_lm$fitted.values,m\_lm$residuals)



qplot(y=m\_lm$fitted.values,x=m\_lm$residuals)



**Вывод о случайном характере остатков:** судя по графику остатки носят случайный характер, так как на графике нет направленности в расположении точек E.

mean(m\_lm$residuals)

## [1] 4.396576e-16

a<-mean(m\_lm$residuals)  
b<-sd(m\_lm$residuals,na.rm = FALSE)  
n<-sqrt(110)  
tm<-a/b\*n  
str(tm)

## num 1.22e-15

qt(0.95,42125)

## [1] 1.64489

**Вывод о равенстве матожидания остатков нулю**: так как |tрасч| < tтабл, то принимается нулевая гипотеза о равенстве матожидания остатков нулю.

bptest(m\_lm)

##   
## studentized Breusch-Pagan test  
##   
## data: m\_lm  
## BP = 63.38, df = 4, p-value = 5.645e-13

**Вывод о гомоскедастичности остатков**: так как p-value < 0.05, то гипотеза о гомоскедастичности остатков отвергается.

cor.test(train$Life.expectancy,m\_lm$residuals,method="spearman")

## Warning in cor.test.default(train$Life.expectancy, m\_lm$residuals, method =  
## "spearman"): Cannot compute exact p-value with ties

##   
## Spearman's rank correlation rho  
##   
## data: train$Life.expectancy and m\_lm$residuals  
## S = 7784800, p-value = 3.234e-16  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## 0.3828667

**Вывод о гомоскедастичности остатков**: так как p-value < 0.05, то гипотеза о гомоскедастичности остатков отвергается.

bgtest(m\_lm)

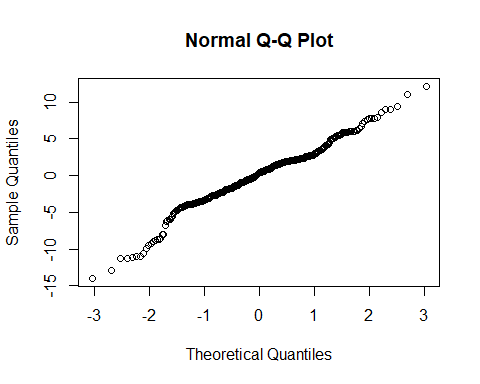
##   
## Breusch-Godfrey test for serial correlation of order up to 1  
##   
## data: m\_lm  
## LM test = 155.24, df = 1, p-value < 2.2e-16

dwtest(m\_lm)

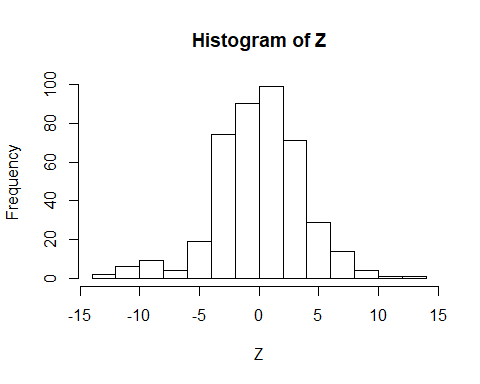
##   
## Durbin-Watson test  
##   
## data: m\_lm  
## DW = 0.85433, p-value < 2.2e-16  
## alternative hypothesis: true autocorrelation is greater than 0

**Вывод**: поскольку значение p-value < 0.05 нулевая гипотеза об отсутствии автокорреляции остатков отвергается.

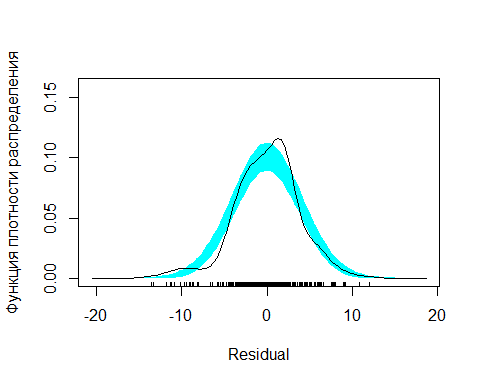
qqnorm(m\_lm$residuals)



library(sm)  
Z<-m\_lm$residuals  
hist(Z)



sm.density(m\_lm$residuals,model="Normal", xlab="Residual",ylab="Функция плотности распределения")

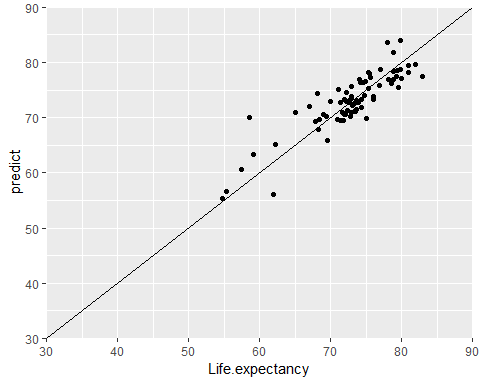


library(nortest)  
lillie.test(m\_lm$residuals)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: m\_lm$residuals  
## D = 0.068421, p-value = 6.367e-05

**Вывод**: так как значение p-value меньше 0.05, нулевая гипотеза о согласии распределения остатков с нормальным законом распределения отвергается.

ggplot(test, aes(x = Life.expectancy, y = predict)) + geom\_point() + geom\_abline() +  
scale\_x\_continuous(limits = c(30, 90), expand = c(0, 0)) +  
scale\_y\_continuous(limits = c(30, 90), expand = c(0, 0))



VIF(m\_lm)

## Adult.Mortality HIV.AIDS   
## 2.546127 1.805178   
## Income.composition.of.resources Schooling   
## 3.249698 3.145869

**Вывод**: отсутствие мультиколлинеарности.

lambdas <- seq(0, 150, by = 0.025)  
set.seed(1877)  
library(glmnet)

## Warning: package 'glmnet' was built under R version 3.5.3

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 3.5.3

## Loading required package: foreach

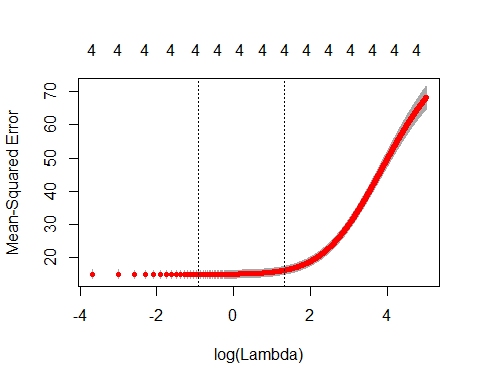
## Warning: package 'foreach' was built under R version 3.5.3

##   
## Attaching package: 'foreach'

## The following object is masked from 'package:memisc':  
##   
## foreach

## Loaded glmnet 2.0-18

cv\_ridge <- cv.glmnet(train\_x, train\_y, alpha = 0, lambda = lambdas)  
plot(cv\_ridge)



cv\_ridge$lambda.min

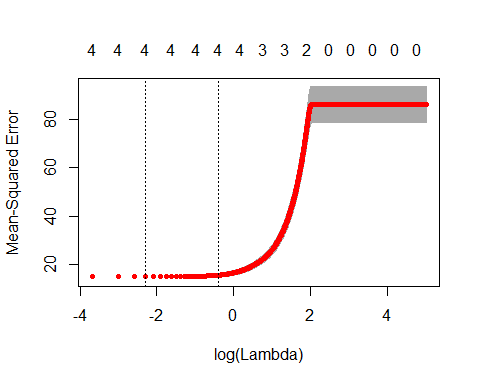
## [1] 0.4

set.seed(1877)  
m\_ridge <- glmnet(train\_x, train\_y, alpha = 0, lambda = cv\_ridge$lambda.min)  
coef(m\_ridge)

## 5 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 56.63054207  
## Adult.Mortality -0.02768886  
## HIV.AIDS -0.36025474  
## Income.composition.of.resources 9.39410141  
## Schooling 1.01885909

test$ridge <- predict(m\_ridge, s = cv\_ridge$lambda.min, newx = test\_x)

set.seed(1886)  
cv\_lasso <- cv.glmnet(train\_x, train\_y, alpha = 1, lambda = lambdas)  
plot(cv\_lasso)



cv\_lasso$lambda.min

## [1] 0.1

set.seed(1886)  
m\_lasso <- glmnet(train\_x, train\_y, alpha = 1, lambda = cv\_lasso$lambda.min)  
coef(m\_lasso)

## 5 x 1 sparse Matrix of class "dgCMatrix"  
## s0  
## (Intercept) 56.88945268  
## Adult.Mortality -0.02903048  
## HIV.AIDS -0.32753556  
## Income.composition.of.resources 8.42103255  
## Schooling 1.06310448

test$lasso <- predict(m\_lasso, s = cv\_lasso$lambda.min, newx = test\_x)

matrix(c("","MAPE","RMSE","MAE",  
 "Method MNK",MAPE(x = test$predict, ref = test$Life.expectancy),RMSE(x = test$predict, ref = test$Life.expectancy),MAE(x = test$predict, ref = test$Life.expectancy),  
 "Method Ridge",  
 MAPE(x = test$ridge, ref = test$Life.expectancy),  
 RMSE(x = test$ridge, ref = test$Life.expectancy),  
 MAE(x = test$ridge, ref = test$Life.expectancy),  
 "Method Lasso",  
 MAPE(x = test$lasso, ref = test$Life.expectancy),  
 RMSE(x = test$lasso, ref = test$Life.expectancy),  
 MAE(x = test$lasso, ref = test$Life.expectancy)),ncol = 4, byrow = TRUE)

## [,1] [,2] [,3]   
## [1,] "" "MAPE" "RMSE"   
## [2,] "Method MNK" "0.0324645661178478" "2.91439685989672"  
## [3,] "Method Ridge" "0.0323490750153504" "2.86905265556552"  
## [4,] "Method Lasso" "0.0326222395344692" "2.91611541007401"  
## [,4]   
## [1,] "MAE"   
## [2,] "2.30671361177805"  
## [3,] "2.29705406993182"  
## [4,] "2.31736503437622"

**Результаты сравнения метрик**: лучше всего показал себя метод Ридж. Они имеют минимальные значения всех трех показателей.

**Вывод о корректировке модели:**

1. Остатки не гомоскедастичны.

2. В нашей модели можно наблюдать автокорреляцию в остатках. То есть имеется зависимость случайных компонент наблюдений с различными номерами. Это приводит к получению неэффективных оценок.

3. Наши остатки не распределенные нормально и это приводит к получению несостоятельных оценок, и как следствие приводящих к ненадежным прогнозам.