r_code

11/30/2021

China Fertility Dataset

import package

```
# load packages
library(olsrr)
```

```
##
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':
##
##     rivers
```

```
library(glmnet)
```

```
## Loading required package: Matrix

## Loaded glmnet 4.1-3
```

```
library(dplyr)
```

```
##
## 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
```

```
library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
library(pls)
```

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

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

```
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

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

```
library(leaps)
library(MASS)
```

```
##
## Attaching package: 'MASS'

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

## The following object is masked from 'package:olsrr':
##
##     cement
```

import dataset

```
Fertility <- read.csv('/Users/karen/Desktop/664/project/dataset.csv', header=T)
```

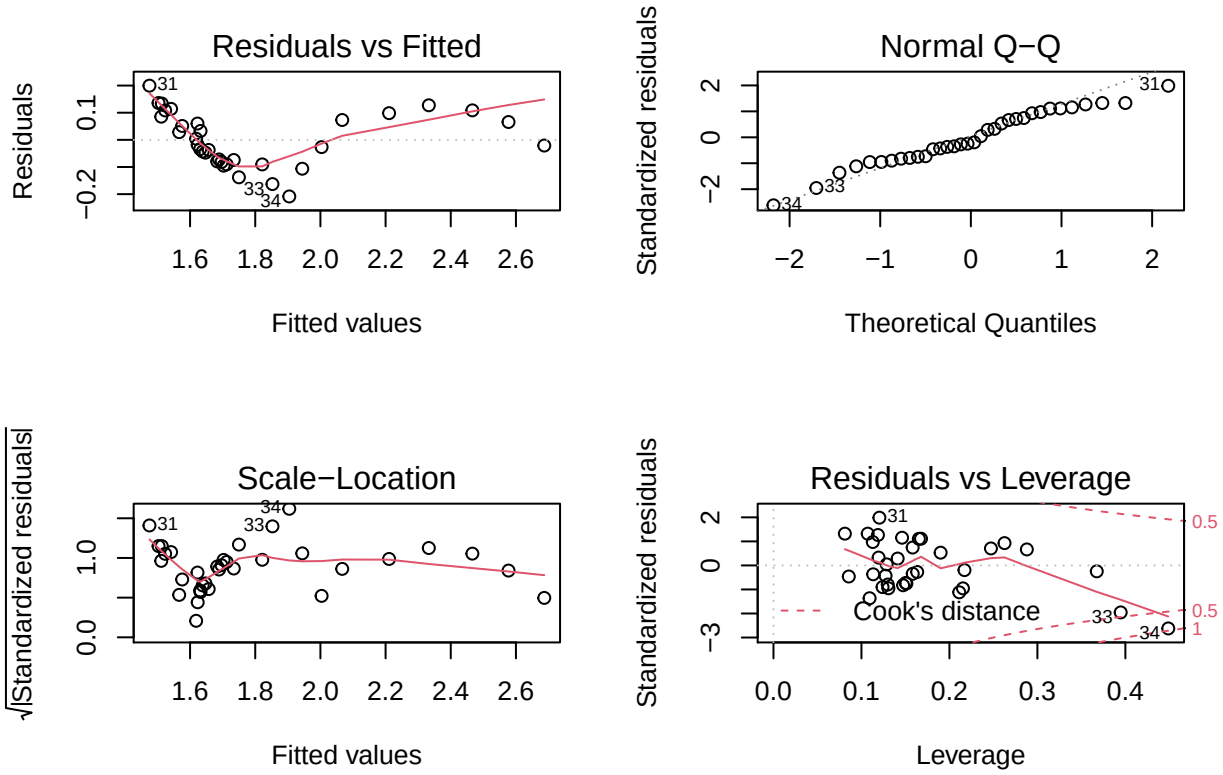
linear regression on the full model

```
lmfull <- lm(fertility~lifexpect+univ.recruit.+urb.popul.+inf.mortality+income.CPI, data = Fertility)
summary(lmfull)
```

```
##
## Call:
## lm(formula = fertility ~ lifexpect + univ.recruit. + urb.popul. +
##     inf.mortality + income.CPI, data = Fertility)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.20849 -0.07811 -0.01973  0.08228  0.19955
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   11.0061914  5.9035562   1.864  0.07279 .
## lifexpect     -0.0292773  0.0738950  -0.396  0.69496
## univ.recruit.  0.0016451  0.0005336   3.083  0.00457 **
## urb.popul.    -0.1744342  0.0211208  -8.259 5.48e-09 ***
## inf.mortality -0.0583659  0.0259102  -2.253  0.03231 *
## income.CPI     0.2989816  0.0407988   7.328 5.59e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1071 on 28 degrees of freedom
## Multiple R-squared:  0.9119, Adjusted R-squared:  0.8962
## F-statistic: 57.96 on 5 and 28 DF,  p-value: 6.687e-14
```

```
#life expectancy is found to be insignificant
```

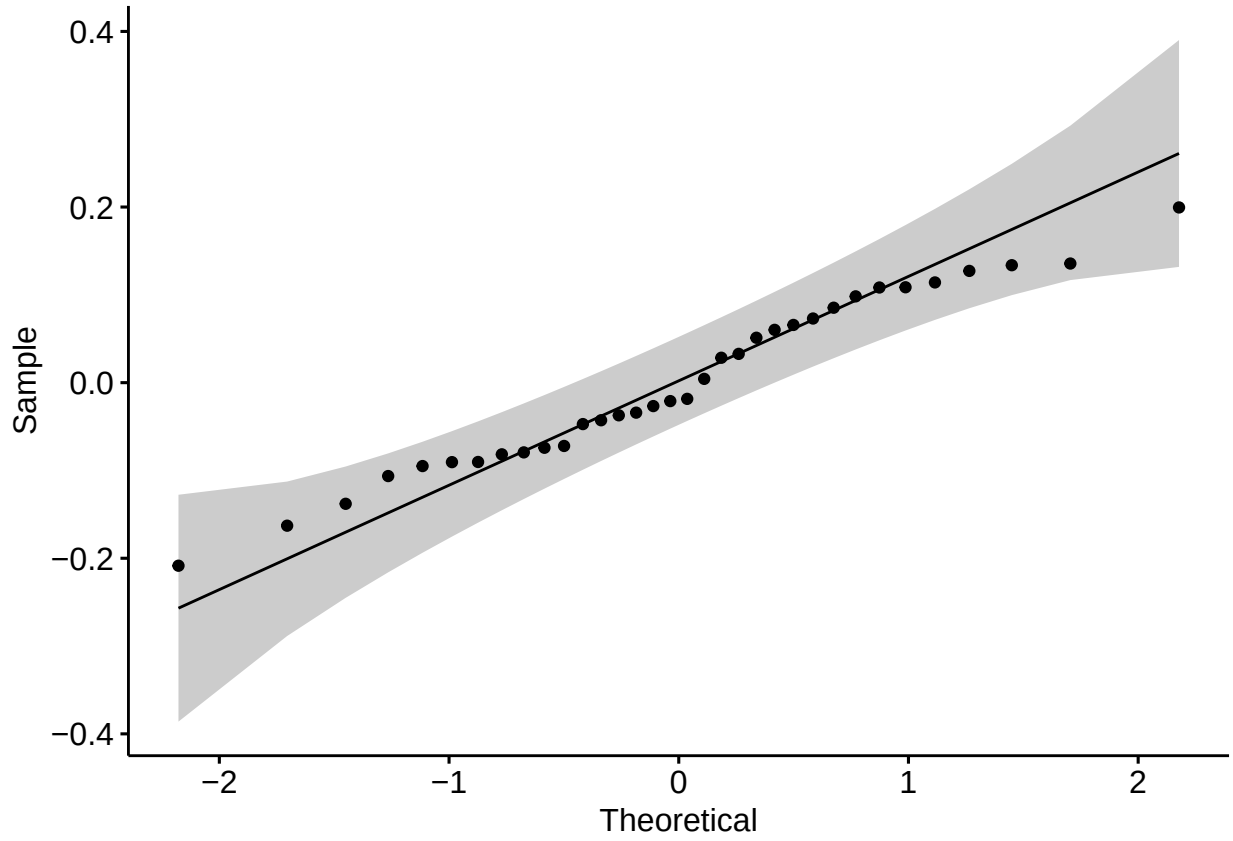
```
par(mfrow=c(2,2))
plot(lmfull)
```



In the Residuals VS Fitted plot, we see that linearity is violated: there seems to be a quadratic relationship. It indicates that there may be a nonlinear relationship between y and xi. There are two outliers,(n=33, n=34) with residuals close to 0.18.

residual normality test

```
library(ggpubr)
ggqqplot(lmfull$residuals)
```




```
shapiro.test(lmfull$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmfull$residuals
## W = 0.97401, p-value = 0.5802
```

```
#p=0.5802>0.05, the normality of residuals is significant
```

Outlier test

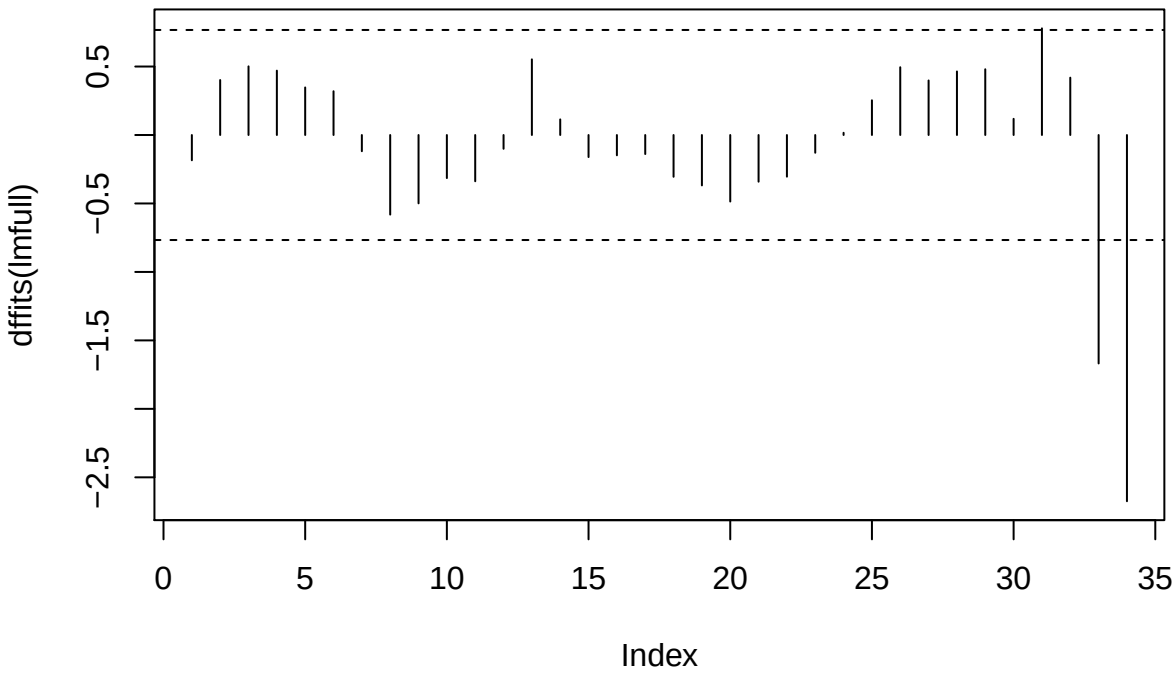
```
library(car)
outlierTest(lmfull)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 34 -2.964366          0.0062691      0.21315
```

```
#DFFITS testing for influential points of the fullmode
dffitsfull<-as.data.frame(dffits(lmfull))
dffitsfull
```

```
##    dffits(lmfull)
## 1     -0.18477374
## 2      0.40155448
## 3      0.50132973
## 4      0.47003770
## 5      0.34729866
## 6      0.31960655
## 7     -0.11840985
## 8     -0.58146538
## 9     -0.49915925
## 10    -0.31454078
## 11    -0.33790760
## 12    -0.10083069
## 13     0.55252790
## 14     0.11399456
## 15    -0.16113009
## 16    -0.14835182
## 17    -0.13903615
## 18    -0.30516350
## 19    -0.36817686
## 20    -0.48594356
## 21    -0.34111583
## 22    -0.30450605
## 23    -0.12986030
## 24     0.01629301
## 25     0.25361932
## 26     0.49493832
## 27     0.39834288
## 28     0.46426924
## 29     0.48012867
## 30     0.11821593
## 31     0.77884731
## 32     0.41886640
## 33    -1.66842949
## 34    -2.67465361
```

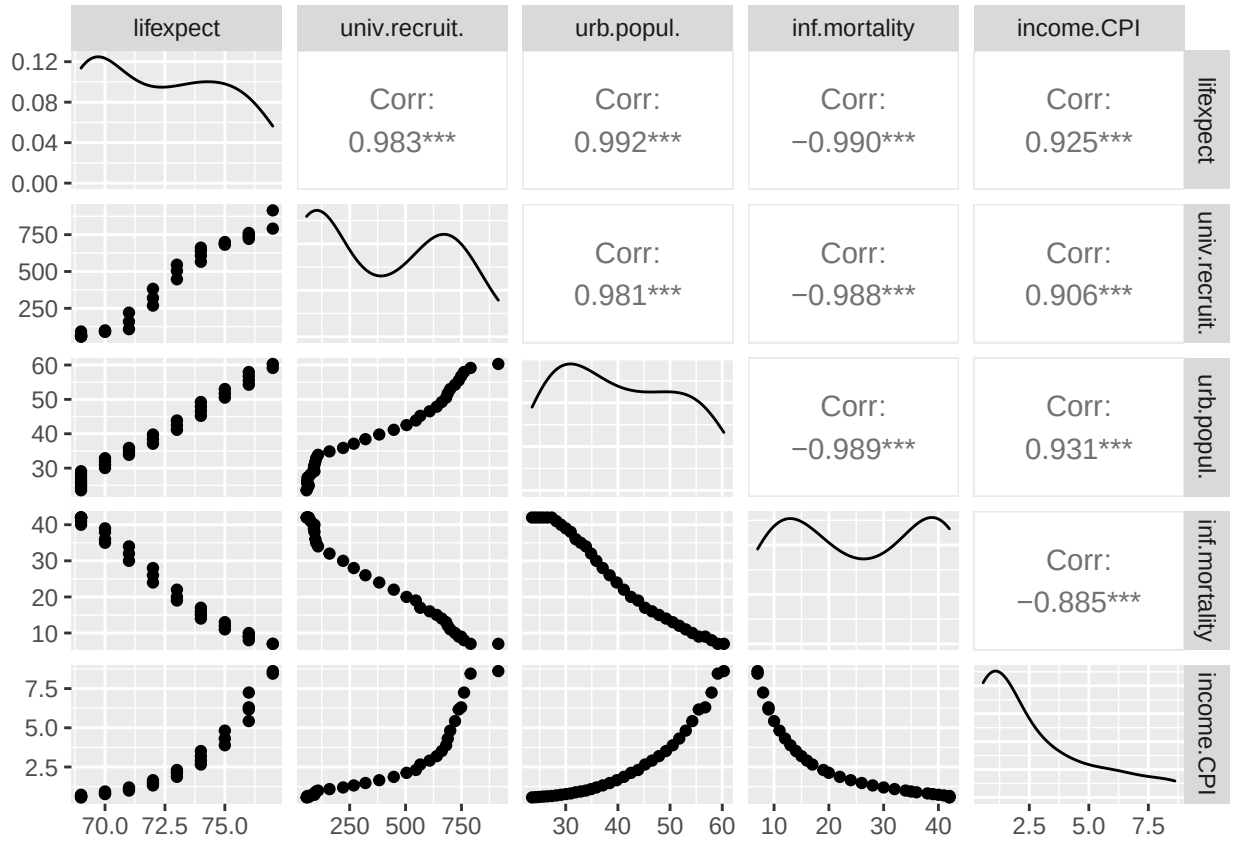
```r
thresholdfull<-2*sqrt(5/34) #p=5, n=34 for fullmode
plot(dffits(lmfull), type = 'h')
abline(h = thresholdfull, lty = 2)
abline(h = -thresholdfull, lty = 2)
```




```
#based on outlier and high-influential test, No. 33 and No. 34 should be excluded from the model

##pairwise correlation
X<-Fertility[,3:7]
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggpairs(X)
```



remove outliers and create a new dataset

```
Fertilitynew <- Fertility[-c(33,34),]
```

model selection

```
#training vs testing slpitting ratio = 5:1
ind <- seq(6, nrow(Fertilitynew), by=6) # an indicating vector from 6 to 30 by 6
fertilitytest <- Fertilitynew[c(ind),]
fertilitytrain <- Fertilitynew[-c(ind),]
fertilitytestx <- fertilitytest[,c(-2,-1)]
fertilitytrainx <- fertilitytrain[,c(-2,-1)]
```

full model

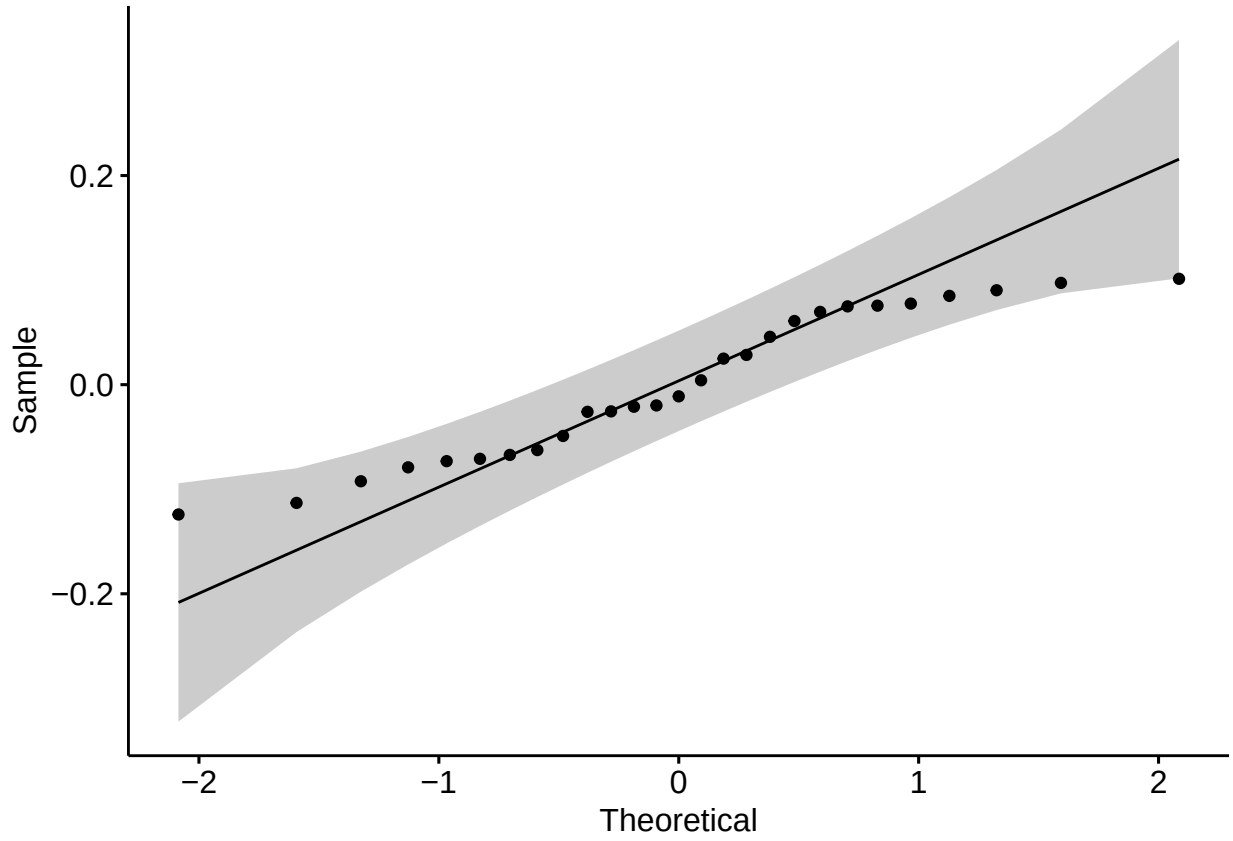
```
lmf <- lm(fertility~lifexpect+univ.recruit.+urb.popul.
          +inf.mortality+income.CPI, data = fertilitytrain)
summary(lmf)
```

```
##
## Call:
## lm(formula = fertility ~ lifexpect + univ.recruit. + urb.popul. +
##     inf.mortality + income.CPI, data = fertilitytrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.12416 -0.06489 -0.01119  0.07223  0.10122
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.2907773  5.1361842   1.030  0.31468
## lifexpect      0.0452860  0.0626363   0.723  0.47766
## univ.recruit.  0.0018593  0.0005288   3.516  0.00205 **
## urb.popul.    -0.1833767  0.0176837 -10.370 1.02e-09 ***
## inf.mortality -0.0398786  0.0246683  -1.617  0.12089
## income.CPI     0.3597430  0.0364981   9.856 2.49e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07898 on 21 degrees of freedom
## Multiple R-squared:  0.9611, Adjusted R-squared:  0.9519
## F-statistic: 103.9 on 5 and 21 DF,  p-value: 4.439e-14
```

```
fertilitytestx <- data.frame(fertilitytestx) # data frame required
#test residual normality
shapiro.test(lmf$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmf$residuals
## W = 0.93121, p-value = 0.07398
```

```
ggqqplot(lmf$residuals)
```



backward selection:

```
#backward selection
lmback1 <- update(lmf,.~.-inf.mortality) #remove infant mortality
summary(lmback1)
```

```
##
## Call:
## lm(formula = fertility ~ lifexpect + univ.recruit. + urb.popul. +
##     income.CPI, data = fertilitytrain)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.168572 -0.057604 -0.003352  0.072630  0.099541
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.2308774  3.2934509  -0.374   0.7122
## lifexpect      0.1075670  0.0511657   2.102   0.0472 *
## univ.recruit.  0.0025712  0.0003032   8.479 2.22e-08 ***
## urb.popul.    -0.1622925  0.0123721 -13.118 7.06e-12 ***
## income.CPI     0.3143821  0.0241811  13.001 8.41e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.08183 on 22 degrees of freedom
## Multiple R-squared:  0.9563, Adjusted R-squared:  0.9484
## F-statistic: 120.4 on 4 and 22 DF,  p-value: 1.277e-14
```

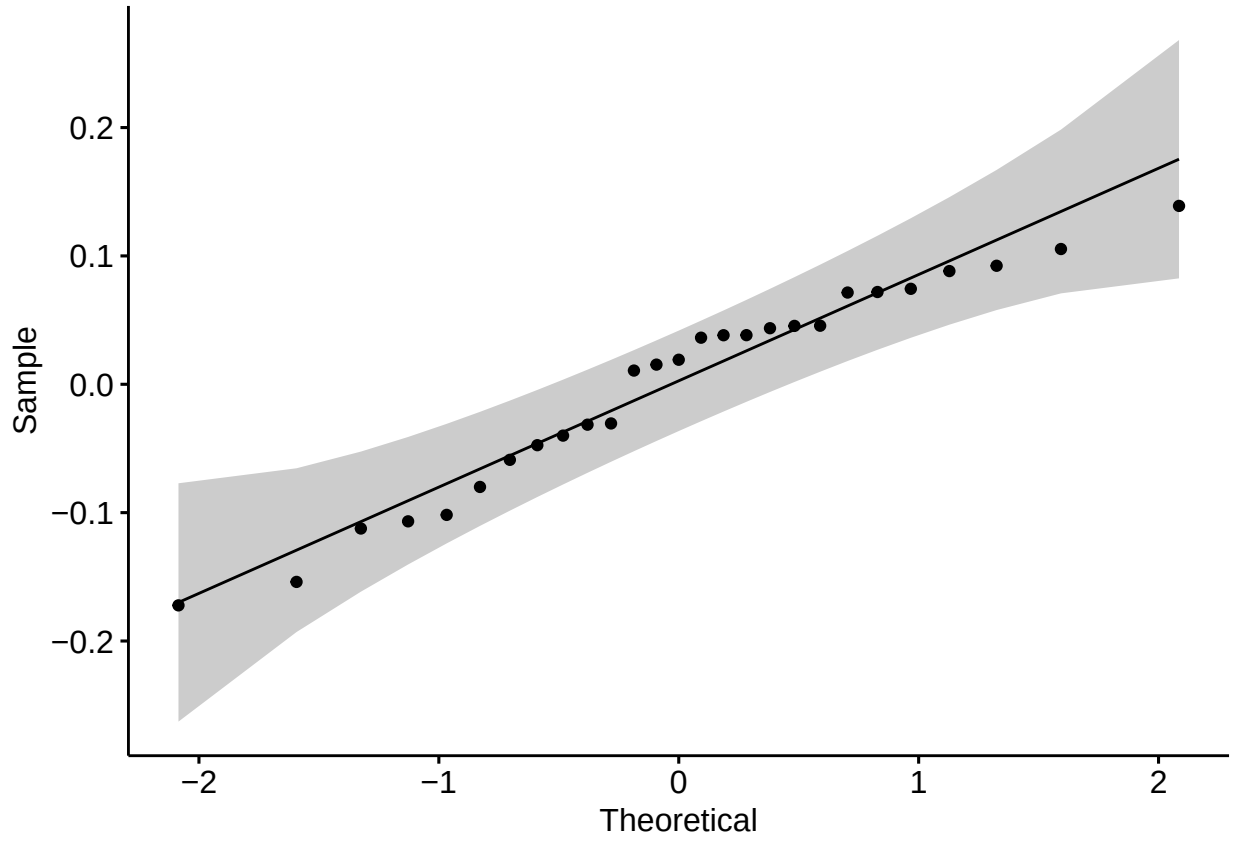
```
lmback2 <- update(lmback1,.~.-lifexpect) # remove life expectancy
summary(lmback2)
```

```
##
## Call:
## lm(formula = fertility ~ univ.recruit. + urb.popul. + income.CPI,
##     data = fertilitytrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.17223 -0.05315  0.01912  0.05854  0.13896
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.6792204  0.2228106  25.489  < 2e-16 ***
## univ.recruit.  0.0028328  0.0002964   9.558 1.78e-09 ***
## urb.popul.    -0.1427480  0.0087494 -16.315 3.88e-14 ***
## income.CPI     0.3073928  0.0256704  11.975 2.31e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0877 on 23 degrees of freedom
## Multiple R-squared:  0.9475, Adjusted R-squared:  0.9407
## F-statistic: 138.4 on 3 and 23 DF,  p-value: 7.329e-15
```

```
#test residual normality
shapiro.test(lmback2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmback2$residuals
## W = 0.95549, p-value = 0.2903
```

```
ggqqplot(lmback2$residuals)
```




```
fertilitytestx <- data.frame(fertilitytestx) # data frame required
#test multicollinearity
car::vif(lmback2)
```

```
## univ.recruit.    urb.popul.    income.CPI
##     22.507023     30.777296      8.345879
```

```
#exist collinearlity
```

AIC

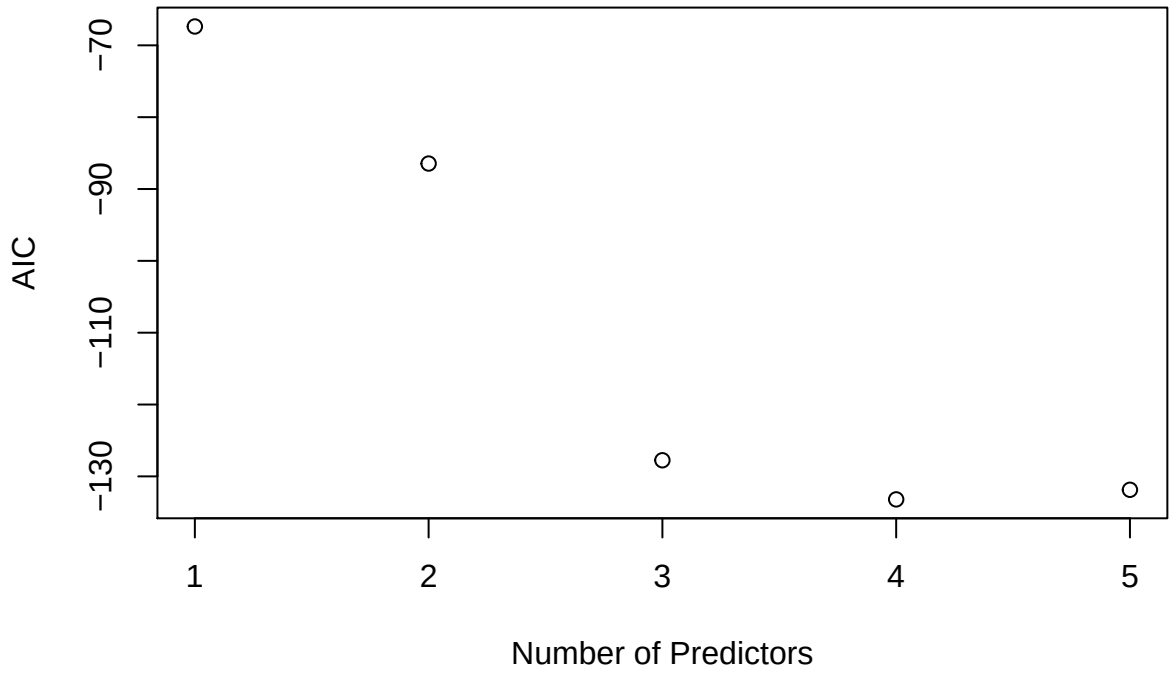
```
aicsub <- regsubsets(fertility~lifexpect+univ.recruit.+urb.popul.
                +inf.mortality+income.CPI, data = fertilitytrain)
rs <- summary(aicsub)
rs$which
```

```
##   (Intercept) lifexpect univ.recruit. urb.popul. inf.mortality income.CPI
## 1        TRUE     FALSE         FALSE       TRUE         FALSE      FALSE
## 2        TRUE     FALSE         FALSE       TRUE         FALSE       TRUE
## 3        TRUE     FALSE          TRUE       TRUE         FALSE       TRUE
## 4        TRUE     FALSE          TRUE       TRUE          TRUE       TRUE
## 5        TRUE      TRUE          TRUE       TRUE          TRUE       TRUE
```

```
rs$rss
```

```
## [1] 1.9199096 0.8795458 0.1769105 0.1342723 0.1310112
```

```
AIC <- 27*log(rs$rss/27) + (2:6)*2 #n=27, p from 2-6
plot(AIC ~ I(1:5), ylab="AIC", xlab="Number of Predictors")
```




```
AIC
```

```
## [1]  -67.37609  -86.45303 -127.75459 -133.20050 -131.86435
```

```
#AIC completed
# all except for lifexpect should be included in this model
lmAIC <- lm(fertility~univ.recruit.+urb.popul.
            +inf.mortality+income.CPI, data = fertilitytrain)
summary(lmAIC)
```

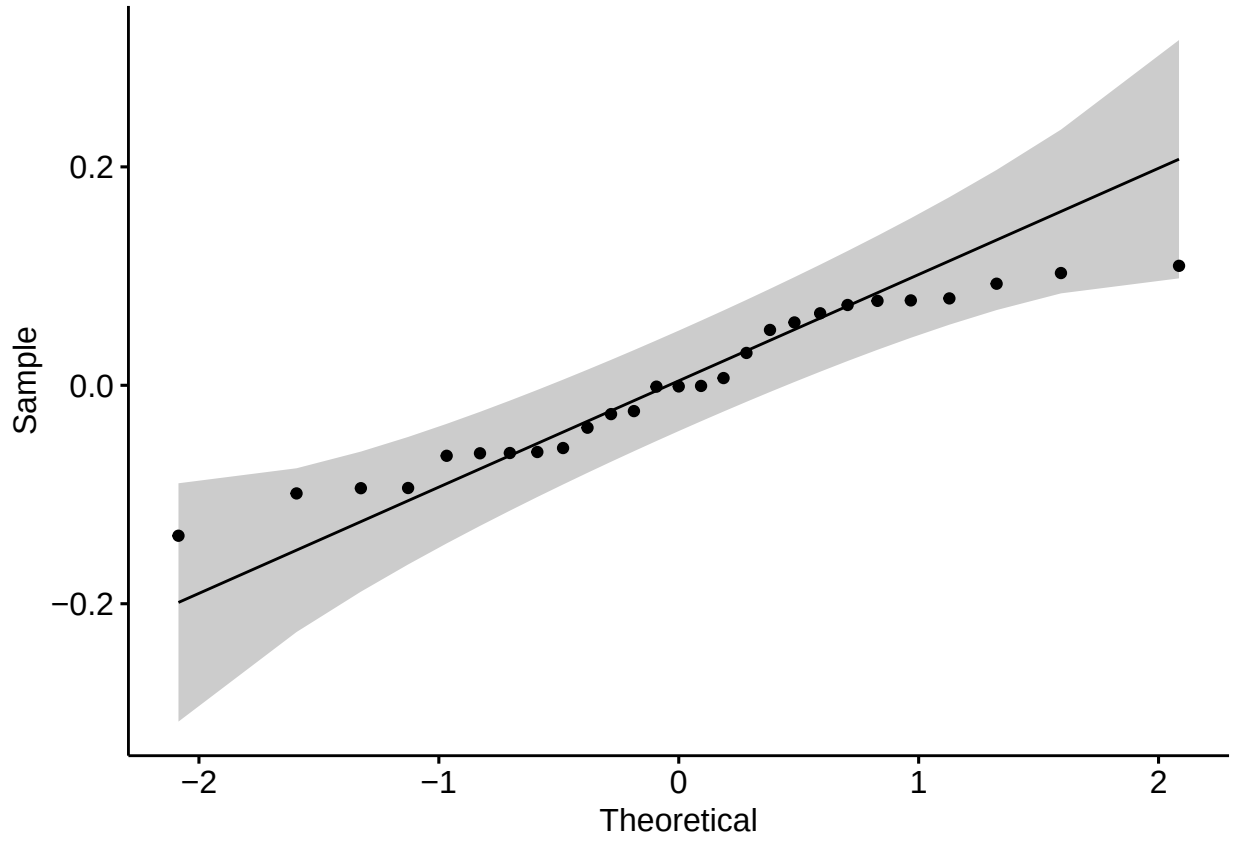
```
##
## Call:
## lm(formula = fertility ~ univ.recruit. + urb.popul. + inf.mortality +
##     income.CPI, data = fertilitytrain)
##
## Residuals:
```

```
##       Min        1Q    Median        3Q       Max
## -0.137742 -0.061530 -0.001025  0.069748  0.109410
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    8.8933550  1.2321259   7.218 3.12e-07 ***
## univ.recruit.  0.0017319  0.0004931   3.512  0.00197 **
## urb.popul.    -0.1840612  0.0174658 -10.538 4.61e-10 ***
## inf.mortality -0.0508485  0.0192380  -2.643  0.01485 *
## income.CPI     0.3703917  0.0330301  11.214 1.44e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07812 on 22 degrees of freedom
## Multiple R-squared:  0.9602, Adjusted R-squared:  0.9529
## F-statistic: 132.6 on 4 and 22 DF,  p-value: 4.622e-15
```

```
#prediction of AIC model
fertilitytestx <- data.frame(fertilitytestx)
# collinearity test
car::vif(lmAIC)
```

```
## univ.recruit.    urb.popul. inf.mortality    income.CPI
##      78.51750     154.56652     248.58715      17.41365
```

```
# normality test
ggqqplot(lmAIC$residuals)
```




```
shapiro.test(lmAIC$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmAIC$residuals
## W = 0.94391, p-value = 0.1521
```

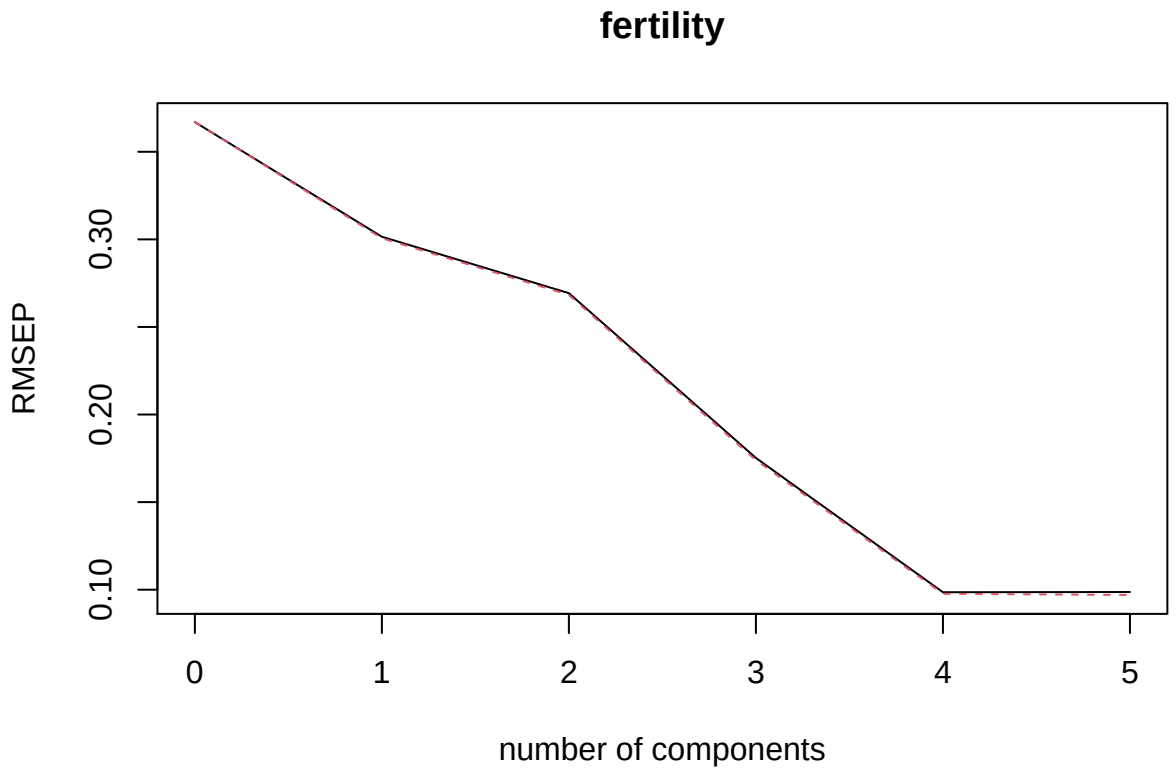
PCR

```
set.seed (1000) # random
lmpcr <- pcr(fertility~lifexpect+univ.recruit.+urb.popul.
             +inf.mortality+income.CPI, data = fertilitytrain,
             scale = TRUE, validation = "CV")
summary(lmpcr)
```

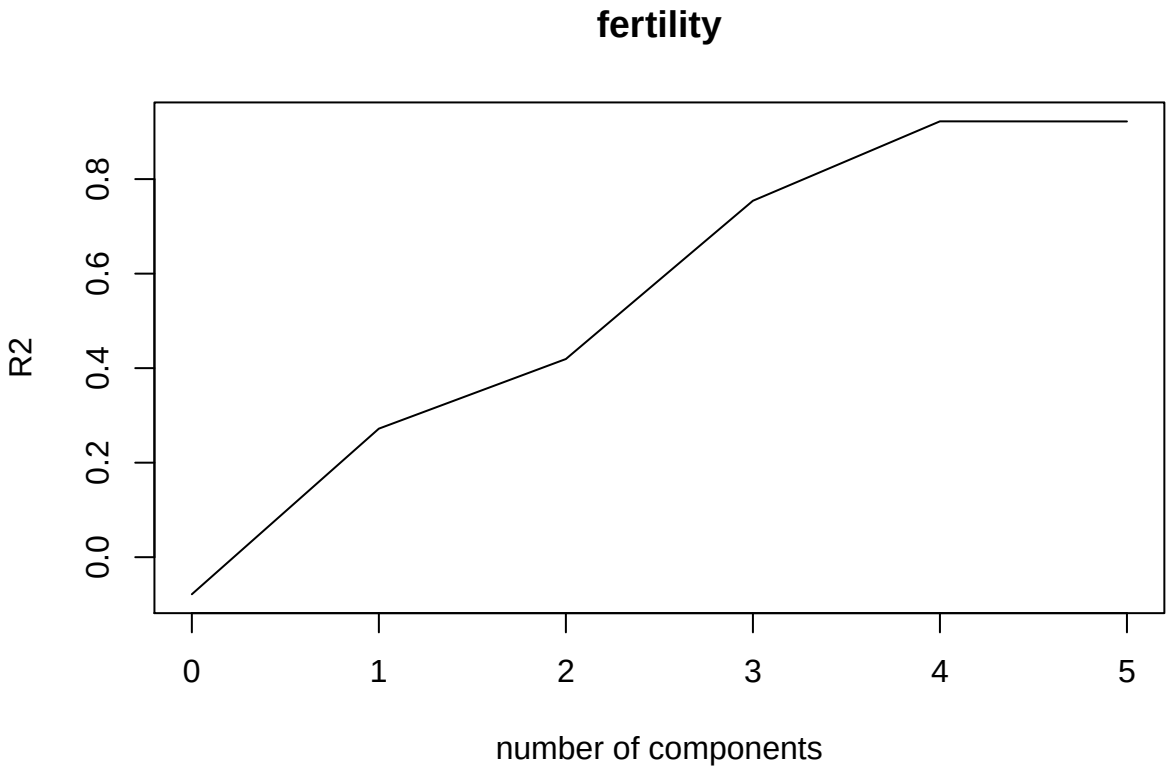
```
## Data:    X dimension: 27 5
##  Y dimension: 27 1
## Fit method: svdpc
## Number of components considered: 5
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
```

```
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV           0.367   0.3015   0.2693   0.1752  0.09855  0.09868
## adjCV        0.367   0.3007   0.2683   0.1740  0.09767  0.09699
##
## TRAINING: % variance explained
##            1 comps  2 comps  3 comps  4 comps  5 comps
## X            96.84    99.32    99.80    99.97   100.00
## fertility    34.62    50.95    82.05    94.52    96.11
```

```
# Plot the root mean squared error
validationplot(lmpcr)
```




```
# Plot the R2
validationplot(lmpcr, val.type = "R2")
```



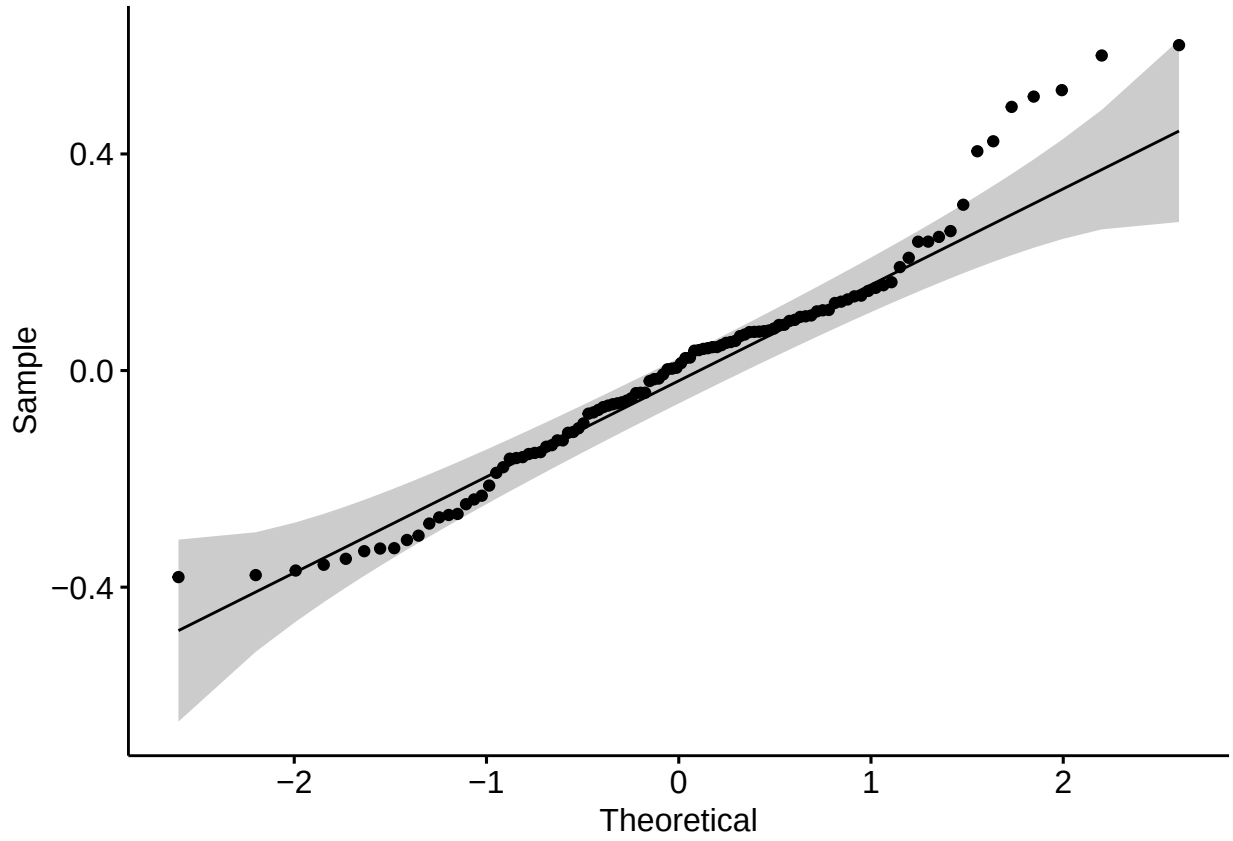

```
# RMSEP test showed an minimum adjusted CV at ncomp=4, while the fifth component contributes little to
# best pcr model
lmpcrbest <- pcr(fertility~lifexpect+univ.recruit.+urb.popul.
            +inf.mortality+income.CPI, data = fertilitytrain,
             scale = TRUE, validation = "CV", ncomp=4)
a <- unlist(lmpcrbest$residuals)
class(a) # list to vector transformation
```

```
## [1] "array"
```

```
shapiro.test(a)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  a
## W = 0.96194, p-value = 0.003526
```

```
ggqqplot(as.numeric(a))
```

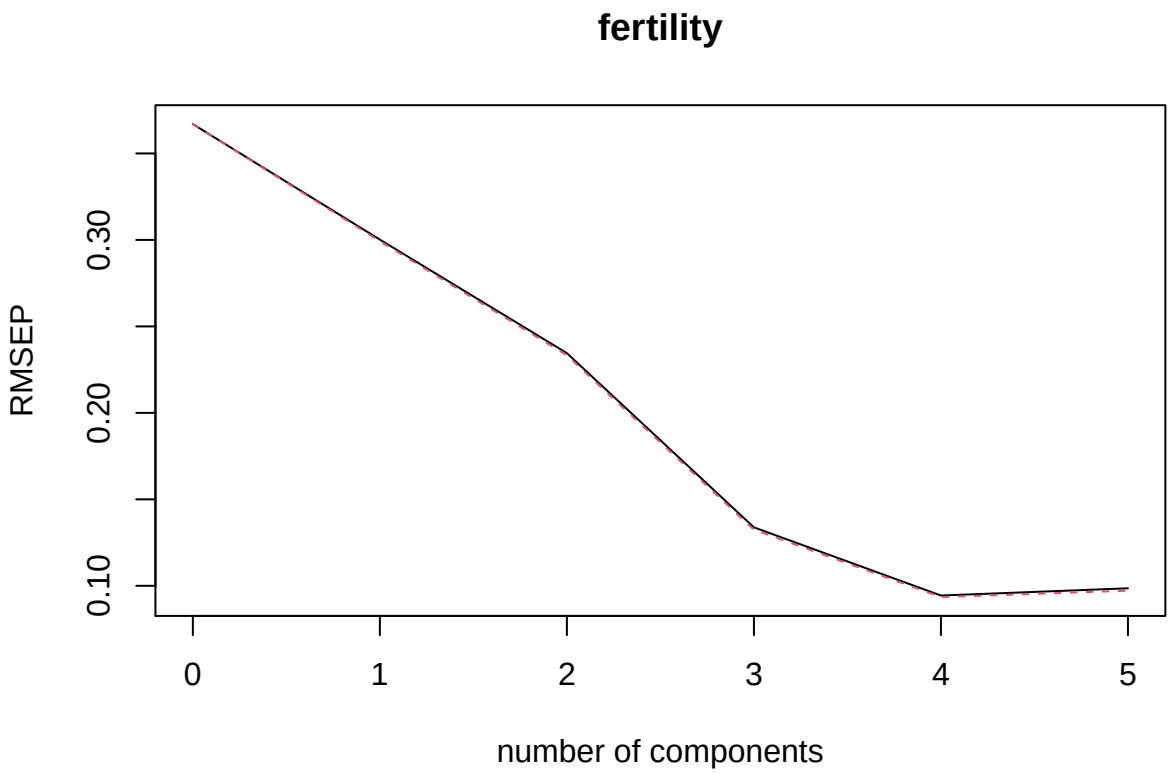


PLSR

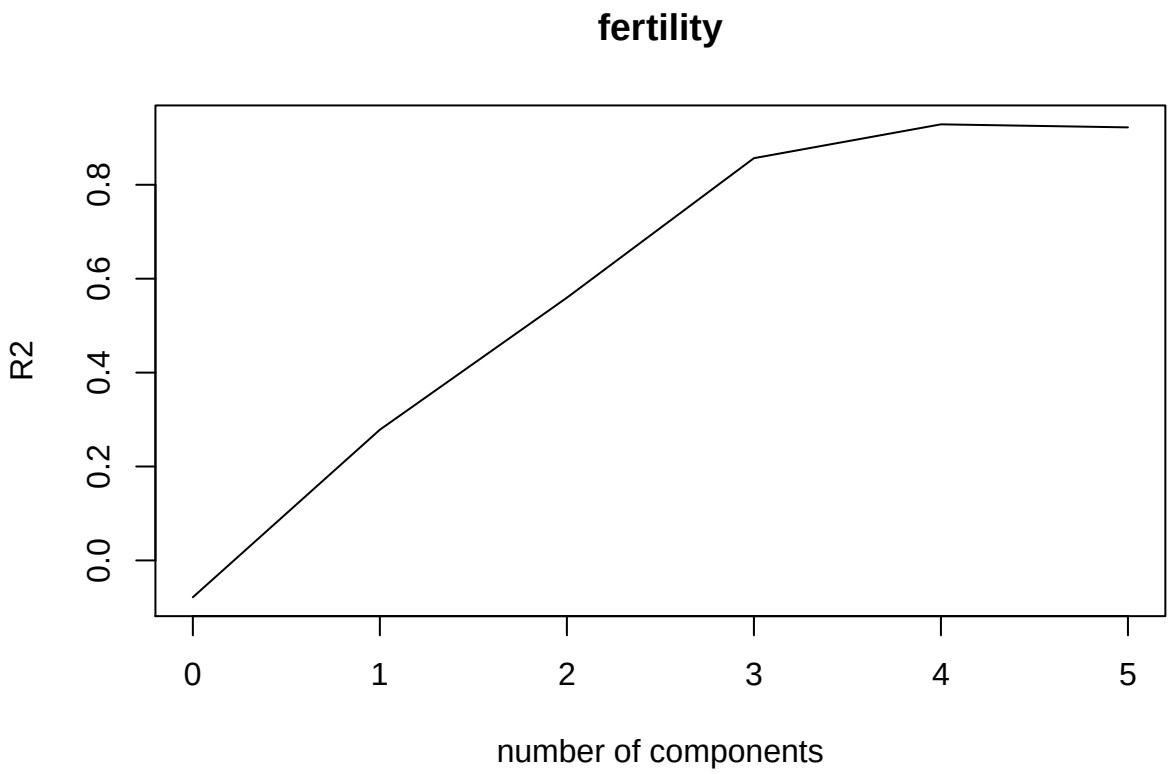
```
set.seed (100)
lmplsr <- plsr(fertility~lifexpect+univ.recruit.+urb.popul.
               +inf.mortality+income.CPI, data = fertilitytrain,
               scale = TRUE, validation = "CV")
summary(lmplsr)
```

```
## Data:    X dimension: 27 5
##  Y dimension: 27 1
## Fit method: kernelpls
## Number of components considered: 5
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps
## CV           0.367   0.3002   0.2345   0.1338  0.09437  0.09856
## adjCV        0.367   0.2993   0.2333   0.1324  0.09352  0.09716
##
## TRAINING: % variance explained
##            1 comps  2 comps  3 comps  4 comps  5 comps
## X            96.81    99.17    99.78    99.97   100.00
## fertility    35.81    65.73    90.51    95.26    96.11
```

```
validationplot(lmplsr)
```




```
validationplot(lmplsr, val.type = "R2")
```



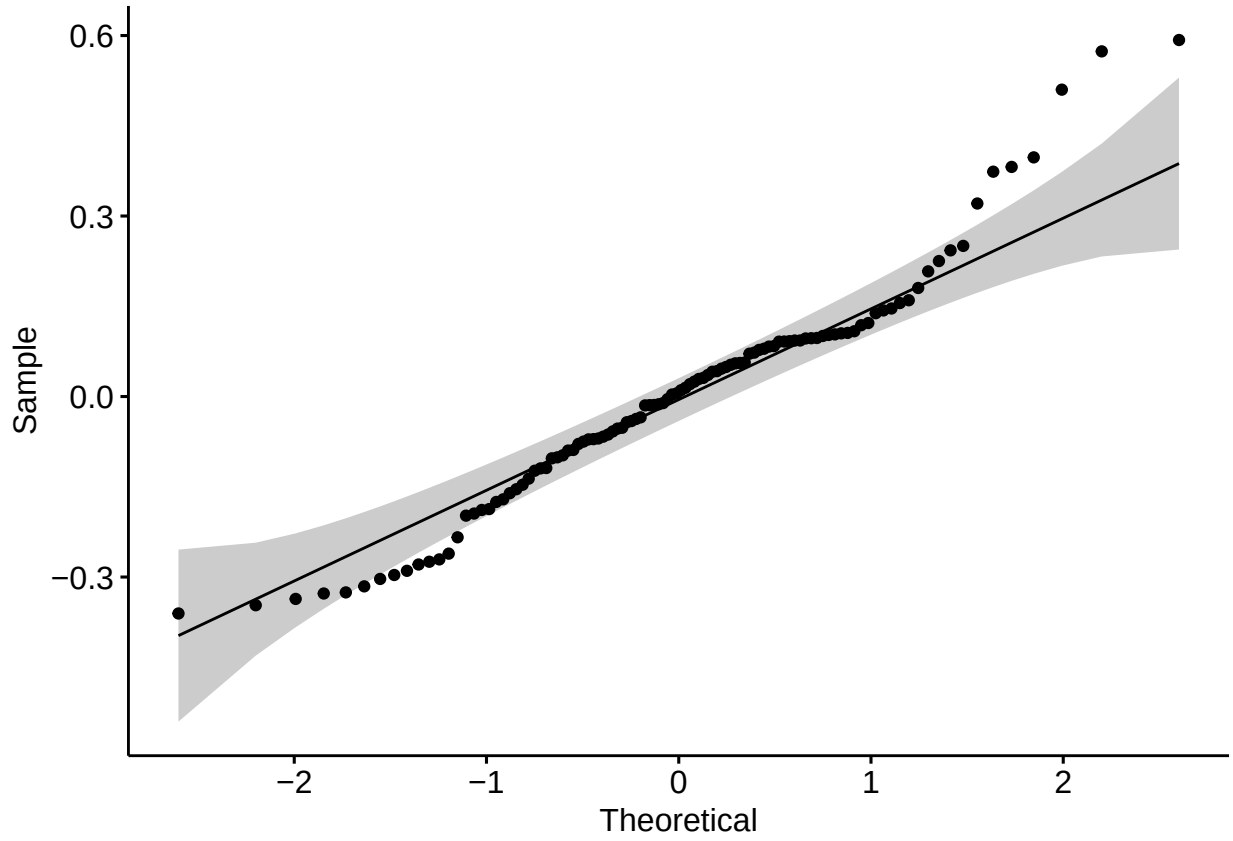

```
# Similar to pcr, ncomp=4 should be enough for the prediction
# best plsr model
lmplsrbest <- plsr(fertility~lifexpect+univ.recruit.+urb.popul.
                +inf.mortality+income.CPI, data = fertilitytrain,
                 scale = TRUE, validation = "CV", ncomp=4)

b <- unlist(lmplsrbest$residuals)
b <- as.numeric(b)
class(b) # list to vector transformation
```

```
## [1] "numeric"
```

```
shapiro.test(b)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  b
## W = 0.95929, p-value = 0.002228
```

```
ggqqplot(b)
```



Ridge regression

```
# Setting the range of lambda values
lambda_seq <- 10^seq(2, -2, by = -.1)
#data transformed to a matrix
fertilitytrainx <- data.matrix(fertilitytrainx)
fertilitytestx <- data.matrix(fertilitytestx)
# build up ridge regression
lmridge <- glmnet(fertilitytrainx, fertilitytrain$fertility,
alpha = 0, lambda  = lambda_seq)
# Checking the model
summary(lmridge)
```

```
##           Length Class     Mode
## a0         41    -none-    numeric
## beta      205    dgCMatrix S4
## df         41    -none-    numeric
## dim         2    -none-    numeric
## lambda     41    -none-    numeric
## dev.ratio  41    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
```

```
## call        5    -none-    call
## nobs        1    -none-    numeric
```

```
# Using cross validation glmnet
ridge_cv <- cv.glmnet(fertilitytrainx, fertilitytrain$fertility,
                      alpha = 0, lambda = lambda_seq)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per
## fold
```

```
# Best lambda value
best_lambda1 <- ridge_cv$lambda.min
best_lambda1
```

```
## [1] 0.01
```

```
# the best lambda 0.01
lmridgebest <- glmnet(fertilitytrainx, fertilitytrain$fertility,
                      alpha = 0, lambda = best_lambda1)
summary(lmridgebest)
```

```
##           Length Class     Mode
## a0        1      -none-    numeric
## beta      5      dgCMatrix S4
## df        1      -none-    numeric
## dim       2      -none-    numeric
## lambda    1      -none-    numeric
## dev.ratio 1      -none-    numeric
## nulldev   1      -none-    numeric
## npasses   1      -none-    numeric
## jerr      1      -none-    numeric
## offset    1      -none-    logical
## call      5      -none-    call
## nobs      1      -none-    numeric
```

```
coef(lmridgebest)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept)    6.255357676
## lifexpect     -0.050229886
## univ.recruit.  0.001180142
## urb.popul.    -0.051292764
## inf.mortality  0.013633051
## income.CPI     0.174868761
```

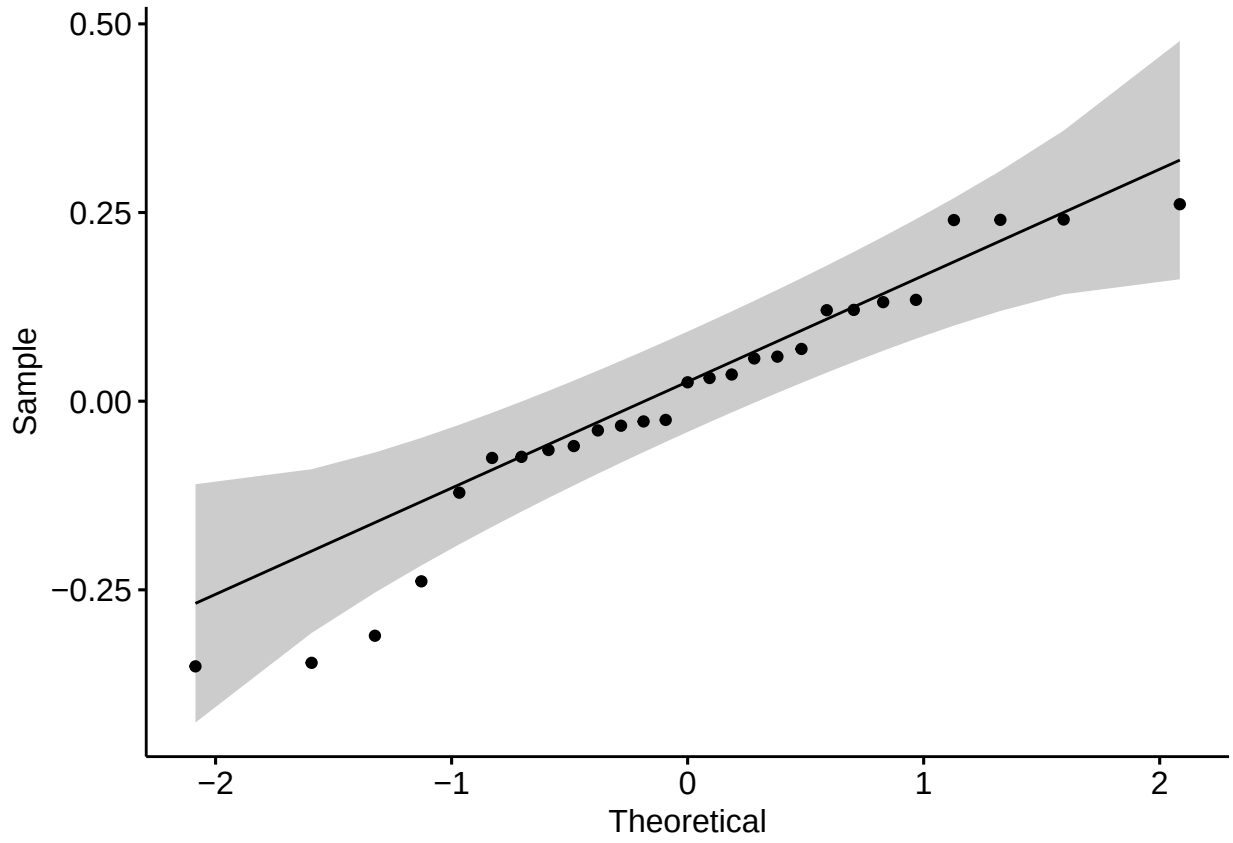
```
fertilitytestx <- data.matrix(fertilitytestx)
# c is the residuals vector
c <- predict(lmridgebest, fertilitytrainx)-fertilitytrain$fertility
c <- as.numeric(c)
class(c) # list to vector transformation
```

```
## [1] "numeric"
```

```
shapiro.test(c)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  c
## W = 0.93875, p-value = 0.1135
```

```
ggqqplot(c)
```




```
# normality is acceptable
```

###LASSO regression

```
# build up LASSO regression
lmlasso <- glmnet(fertilitytrainx, fertilitytrain$fertility, alpha = 1, lambda  = lambda_seq)
# Checking the model
summary(lmlasso)
```

```
##           Length Class     Mode
## a0         41    -none-    numeric
```

```
## beta      205    dgCMatrix S4
## df         41    -none-    numeric
## dim         2    -none-    numeric
## lambda     41    -none-    numeric
## dev.ratio  41    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        5    -none-    call
## nobs        1    -none-    numeric
```

```
# Using cross validation glmnet
lasso_cv <- cv.glmnet(fertilitytrainx, fertilitytrain$fertility, alpha = 1, lambda  = lambda_seq)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per
## fold
```

```
# Best lambda value
best_lambda2 <- lasso_cv$lambda.min
best_lambda2
```

```
## [1] 0.01
```

```
# the best lambda 0.01
lmlassobest <- glmnet(fertilitytrainx, fertilitytrain$fertility, alpha = 1, lambda  = best_lambda2)
summary(lmlassobest)
```

```
##           Length Class     Mode
## a0        1      -none-    numeric
## beta      5      dgCMatrix S4
## df        1      -none-    numeric
## dim       2      -none-    numeric
## lambda    1      -none-    numeric
## dev.ratio 1      -none-    numeric
## nulldev   1      -none-    numeric
## npasses   1      -none-    numeric
## jerr      1      -none-    numeric
## offset    1      -none-    logical
## call      5      -none-    call
## nobs      1      -none-    numeric
```

```
coef(lmlassobest)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)    4.233018051
## lifexpect      .
## univ.recruit.  0.001140735
## urb.popul.    -0.084999078
## inf.mortality  .
## income.CPI     0.216934798
```

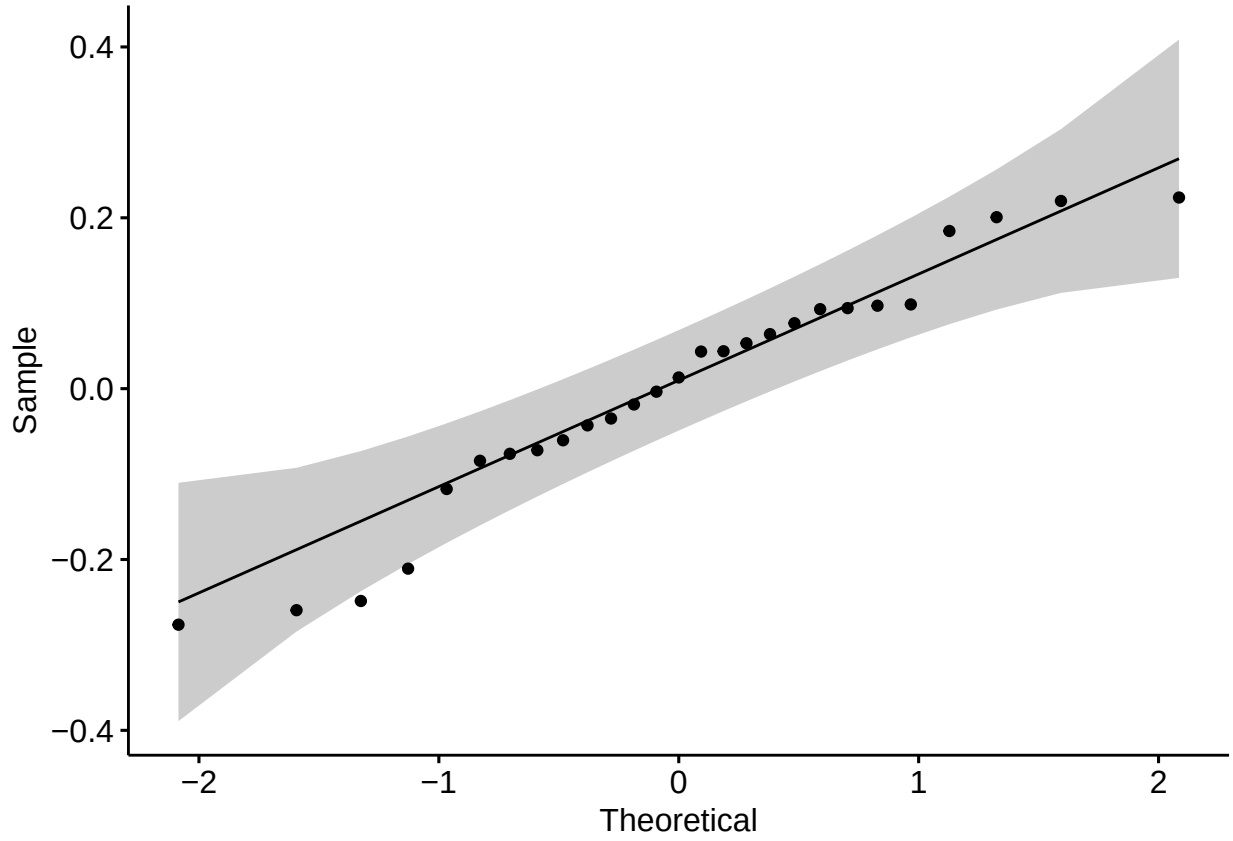
```
fertilitytestx <- data.matrix(fertilitytestx)
# c is the residuals vector
d <- predict(lmlassobest, fertilitytrainx)-fertilitytrain$fertility
d <- as.numeric(d)
class(d) # list to vector transformation
```

```
## [1] "numeric"
```

```
shapiro.test(d)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  d
## W = 0.95354, p-value = 0.2609
```

```
ggqqplot(d)
```




```
# normality is good
```

model prediction performance

```
library(Metrics)
#full model
p_full <- predict(lmf, newdata=fertilitytest)
rmse_full <- rmse(fertilitytest$fertility,p_full)

#backward selection
p_back<-predict(lmback2, fertilitytest)
rmse_back <- rmse(fertilitytest$fertility,p_back)

#AIC
p_aic <- predict.lm(lmAIC, fertilitytest)
rmse_aic <- rmse(fertilitytest$fertility,p_aic)

# ncomp=4, prediction of PCR model
p_pcr <- predict(lmpcr,fertilitytest,ncomp=4)
rmse_pcr <- rmse(fertilitytest$fertility, p_pcr)

# prediction of PLSR model
p_plsr <- predict(lmplsrbest, fertilitytest, ncomp = 4)
rmse_plsr <- rmse(fertilitytest$fertility, p_plsr)

#ridge regression
p_ridge <- predict(lmridgebest, fertilitytestx)
rmse_ridge <- rmse(fertilitytest$fertility, p_ridge)

#LASSO
p_lasso<- predict(lmlassobest, fertilitytestx)
rmse_lasso <- rmse(fertilitytest$fertility, p_lasso)

rmse <- data.frame(method =
                     c('full', 'backward', 'AIC', 'PCR', 'PLS', 'Ridge', 'LASSO'),
                   rmse = c(rmse_full, rmse_back, rmse_aic, rmse_pcr,rmse_plsr, rmse_ridge,rmse_lasso))
rmse
```

```
##     method       rmse
## 1     full 0.06446588
## 2 backward 0.04200628
## 3      AIC 0.06129241
## 4      PCR 0.07551262
## 5      PLS 0.07124705
## 6    Ridge 0.09737348
## 7    LASSO 0.07385271
```

Japan Fertility Dataset

import dataset

```
Fertility2 <- read.csv("/Users/karen/Desktop/664/project/japan_dataset.csv", header=T)
Fertilityx2 <- Fertility2[,-4]
```

linear regression

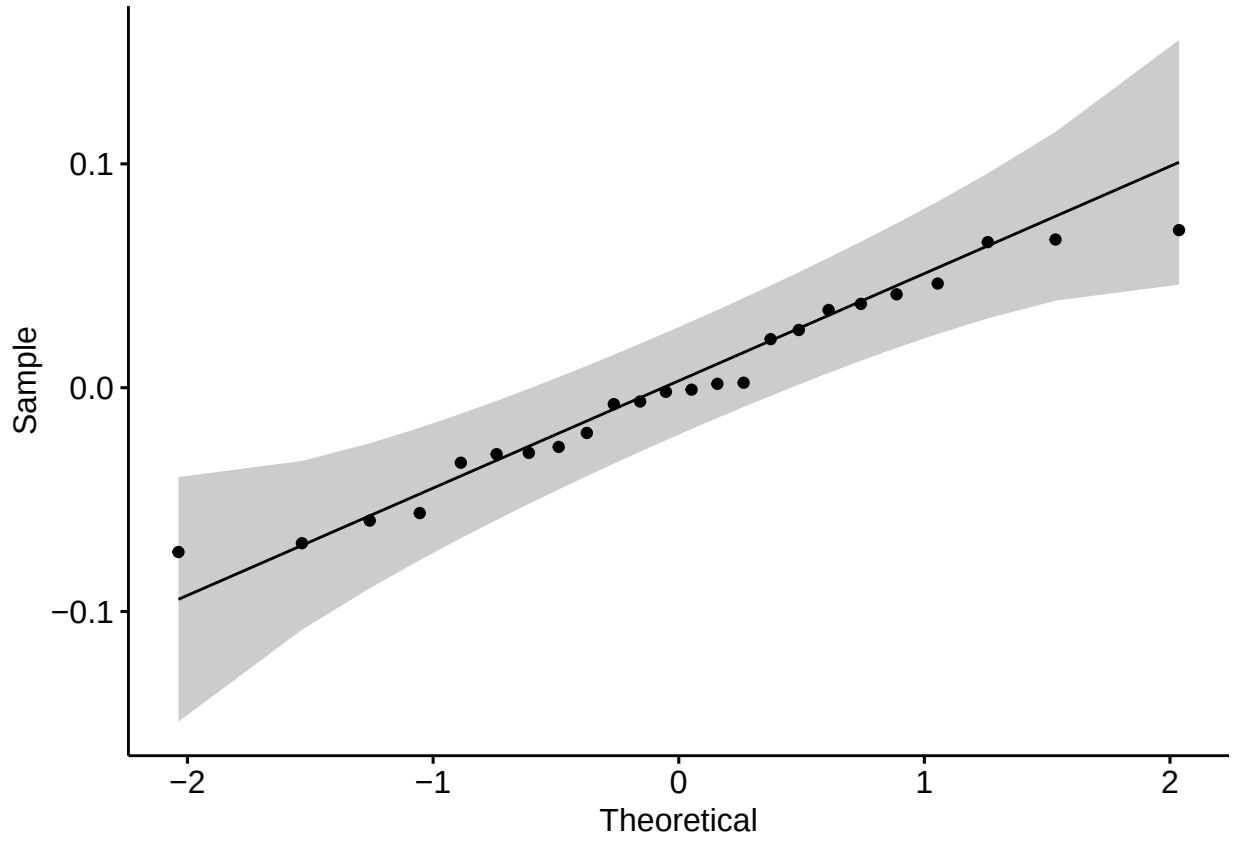
```
lmfull <- lm(fertility~lifexpect+fem.enroll.+urb.popul.
             +inf.mortality+HHDI.CPI, data = Fertility2)
summary(lmfull)
```

```
##
## Call:
## lm(formula = fertility ~ lifexpect + fem.enroll. + urb.popul. +
##     inf.mortality + HHDI.CPI, data = Fertility2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.073418 -0.029254 -0.001356  0.035399  0.070397
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.178016   2.495908  -0.873  0.39435
## lifexpect      0.058227   0.034952   1.666  0.11304
## fem.enroll.   -0.026406   0.007364  -3.586  0.00211 **
## urb.popul.    -0.011531   0.009127  -1.263  0.22256
## inf.mortality  0.009323   0.035209   0.265  0.79418
## HHDI.CPI       0.004635   0.001272   3.644  0.00185 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0484 on 18 degrees of freedom
## Multiple R-squared:  0.504,  Adjusted R-squared:  0.3662
## F-statistic: 3.658 on 5 and 18 DF,  p-value: 0.01851
```

```
#life expectancy is found to be insignificant
```

residual

```
ggqqplot(lmfull$residuals)
```




```
shapiro.test(lmfull$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmfull$residuals
## W = 0.96208, p-value = 0.4817
```

```
#p=0.4817>0.05, the normality of residuals is significant
```

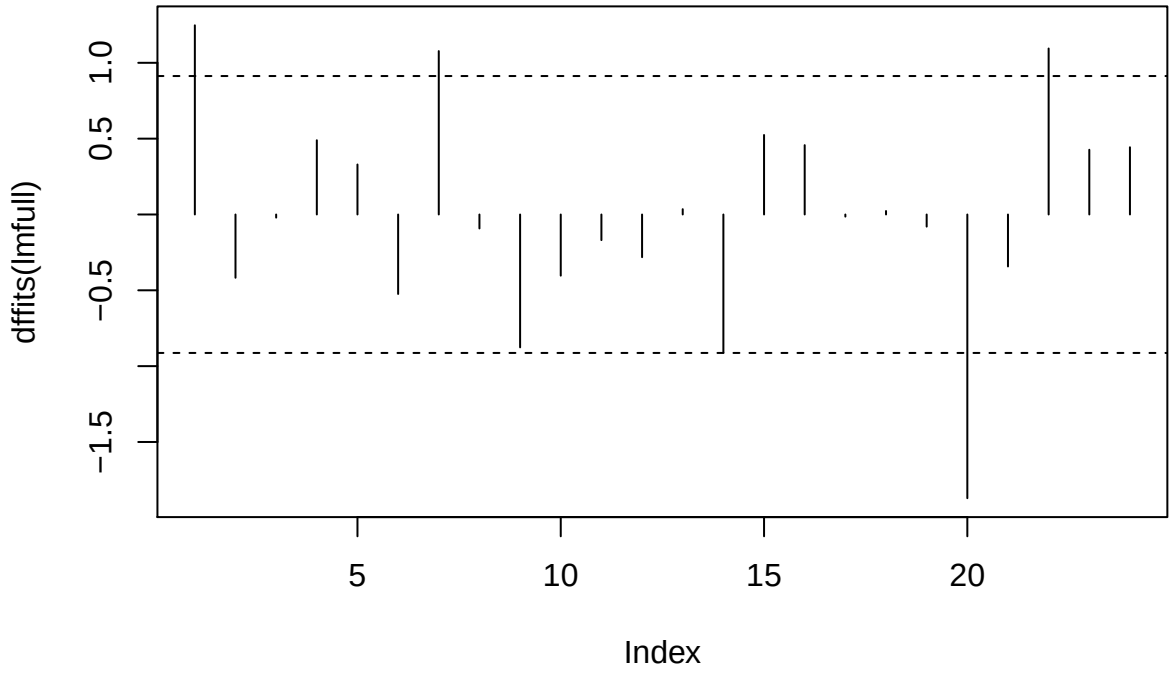
DFFITS testing for influential points of the fullmode

```
dffitsfull<-as.data.frame(dffits(lmfull))
dffitsfull
```

```
##    dffits(lmfull)
## 1      1.24735235
## 2     -0.41730196
## 3     -0.02083207
## 4      0.48953154
## 5      0.32954784
## 6     -0.52411727
## 7      1.07746552
```

```
## 8     -0.09215094
## 9     -0.87616611
## 10    -0.40332356
## 11    -0.16895620
## 12    -0.28160887
## 13     0.03511586
## 14    -0.91180327
## 15     0.52426452
## 16     0.45720424
## 17    -0.01411882
## 18     0.02301465
## 19    -0.07997091
## 20    -1.87058192
## 21    -0.34279923
## 22     1.09445307
## 23     0.42692163
## 24     0.44344291
```

```
thresholdfull<-2*sqrt(5/24) #p=5, n=24 for fullmode
plot(dffits(lmfull), type = 'h')
abline(h = thresholdfull, lty = 2)
abline(h = -thresholdfull, lty = 2)
```




```
#outliers testing with full model
outlierTest(lmfull)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##    rstudent unadjusted p-value Bonferroni p
## 20  -1.8561           0.080859           NA
```

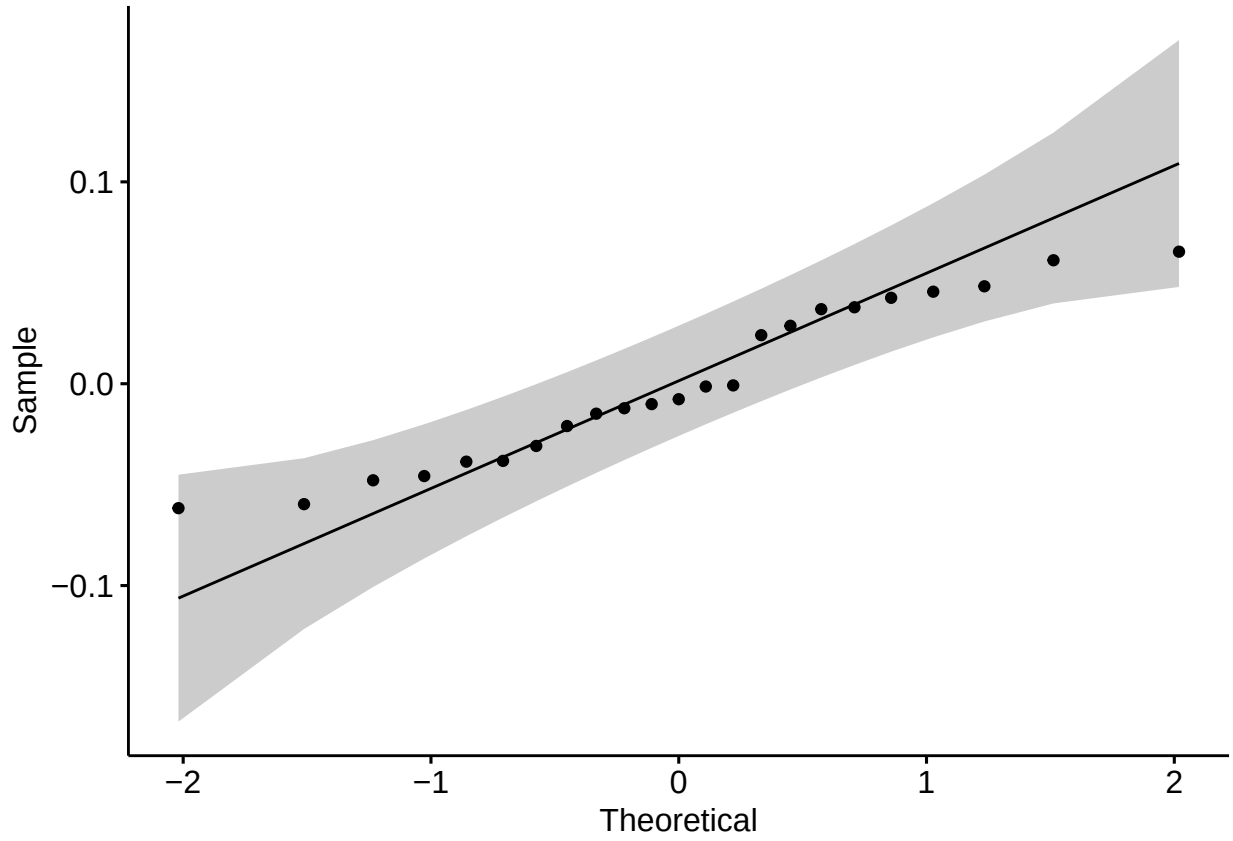
```
# No. 20 is an outlier
Fertilitynew2 <- Fertility2[c(-20),]
Fertilityxnew2 <- Fertilitynew2[,c(-1,-5)]
```

remove outliers and create new model

```
#regression attempt
lmfull1 <- lm(fertility~lifexpect+fem.enroll.+urb.popul.
              +inf.mortality+HHDI.CPI, data = Fertilitynew2)
summary(lmfull1)
```

```
##
## Call:
## lm(formula = fertility ~ lifexpect + fem.enroll. + urb.popul. +
##     inf.mortality + HHDI.CPI, data = Fertilitynew2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.061648 -0.034532 -0.007678  0.037404  0.065408
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -0.324524   2.545924  -0.127 0.900066
## lifexpect      0.038568   0.034464   1.119 0.278673
## fem.enroll.   -0.027688   0.006944  -3.987 0.000953 ***
## urb.popul.    -0.018032   0.009252  -1.949 0.067997 .
## inf.mortality  0.011933   0.033067   0.361 0.722631
## HHDI.CPI       0.006287   0.001489   4.223 0.000572 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04541 on 17 degrees of freedom
## Multiple R-squared:  0.5844, Adjusted R-squared:  0.4621
## F-statistic:  4.78 on 5 and 17 DF,  p-value: 0.00655
```

```
#life expectancy is found to be insignificant
ggqqplot(lmfull1$residuals)
```




```
shapiro.test(lmfull1$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmfull1$residuals
## W = 0.94133, p-value = 0.1919
```

```
#p=0.1919>0.05, the normality of residuals is significant

#multicollinearity test
car::vif(lmfull1)
```

```
##     lifexpect   fem.enroll.    urb.popul. inf.mortality      HHDI.CPI
##     24.090150     38.227406     28.017104      6.685249     53.432223
```

```
# collinearity found
```

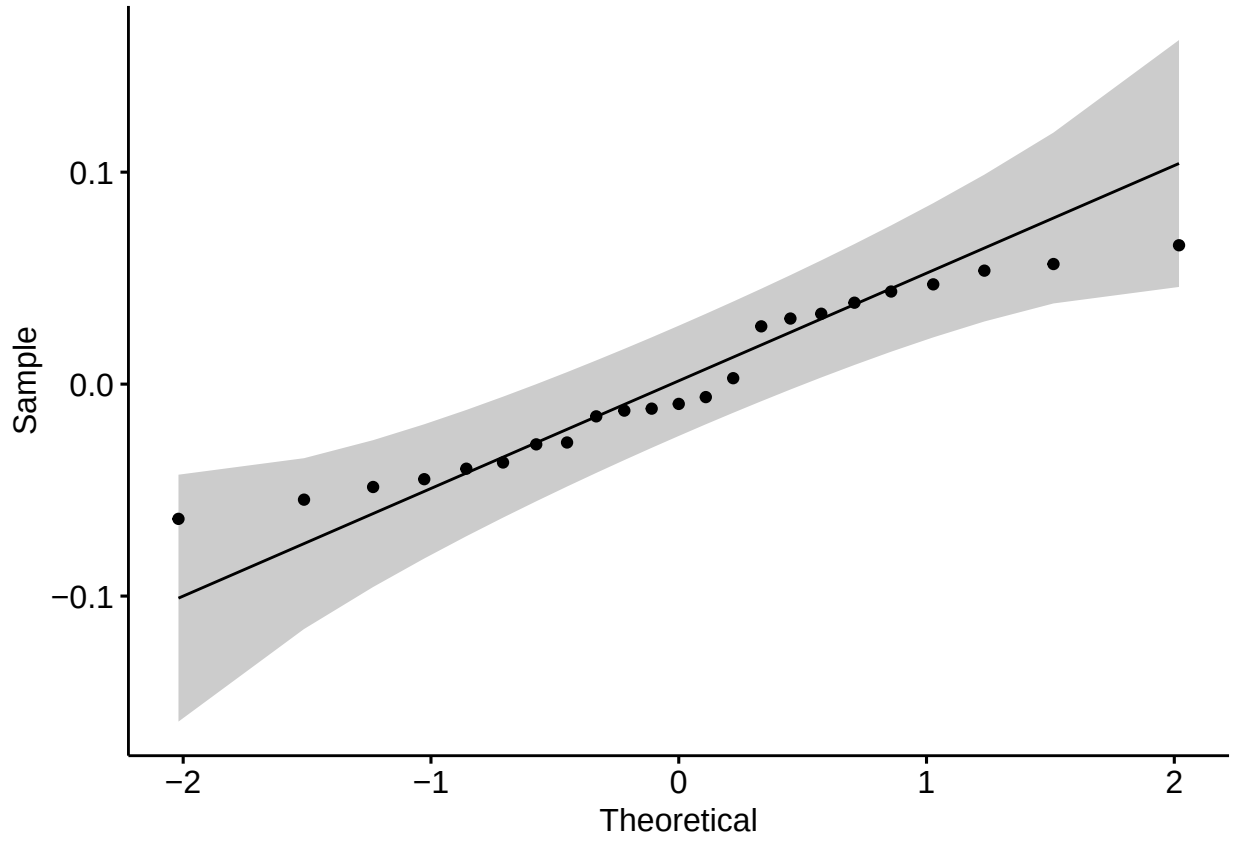
model selection

```
#backward selection
#regression attempt
```

```r
lmback1 <- lm(fertility~lifexpect+fem.enroll.+urb.popul.
              +HHDI.CPI, data = Fertilitynew2)
summary(lmback1)
```

```
##
## Call:
## lm(formula = fertility ~ lifexpect + fem.enroll. + urb.popul. +
##     HHDI.CPI, data = Fertilitynew2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.063573 -0.032696 -0.009334  0.035807  0.065513
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.212665   2.465178  -0.086 0.932206
## lifexpect    0.038075   0.033595   1.133 0.271931
## fem.enroll. -0.028090   0.006686  -4.201 0.000537 ***
## urb.popul.  -0.017980   0.009025  -1.992 0.061725 .
## HHDI.CPI     0.006192   0.001429   4.332 0.000401 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0443 on 18 degrees of freedom
## Multiple R-squared:  0.5812, Adjusted R-squared:  0.4881
## F-statistic: 6.245 on 4 and 18 DF,  p-value: 0.00247
```

```r
#life expectancy is found to be insignificant
ggqqplot(lmback1$residuals)
```




```
shapiro.test(lmback1$residuals)
```

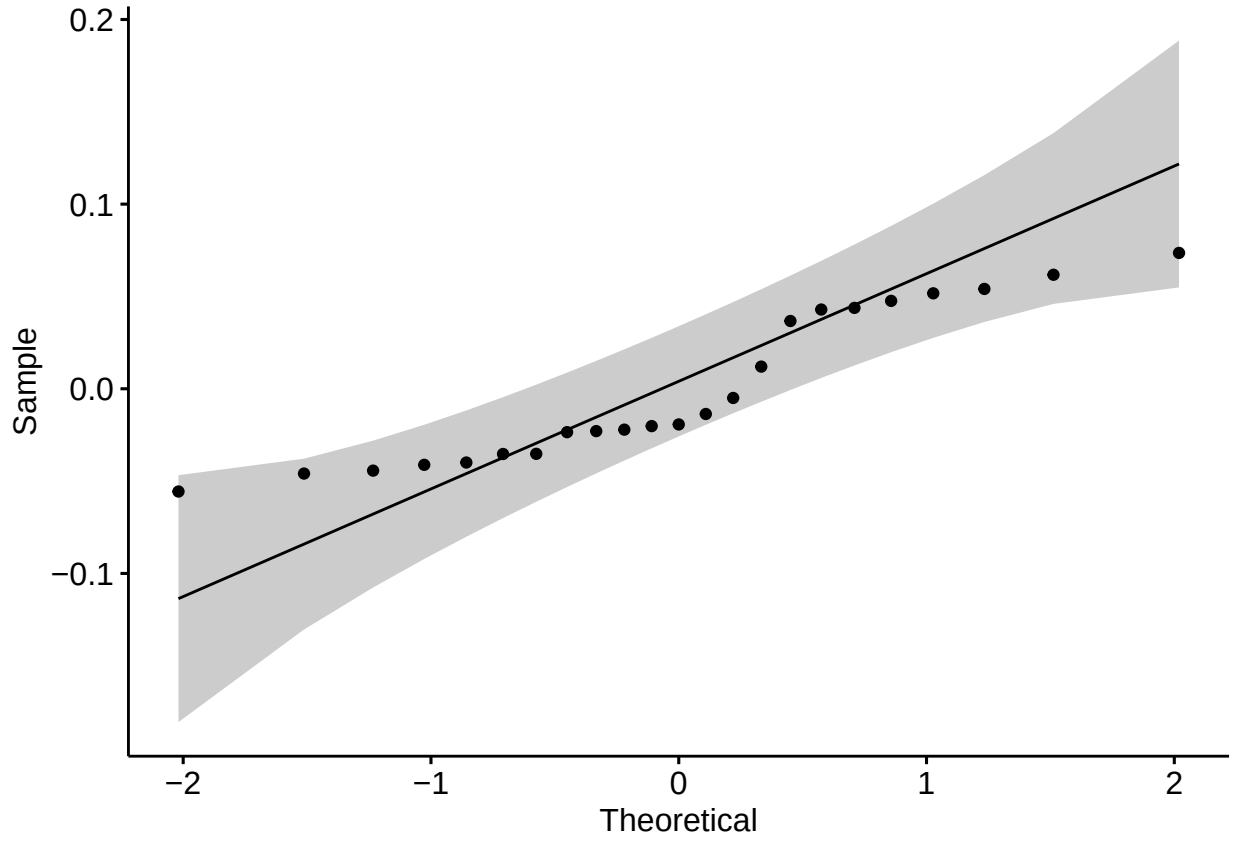
```
##
##  Shapiro-Wilk normality test
##
## data:  lmback1$residuals
## W = 0.93707, p-value = 0.1554
```

```
#p=0.1554>0.05, the normality of residuals is significant

lmback2 <- lm(fertility~fem.enroll.+urb.popul.
              +HHDI.CPI, data = Fertilitynew2)
summary(lmback2)
```

```
##
## Call:
## lm(formula = fertility ~ fem.enroll. + urb.popul. + HHDI.CPI,
##     data = Fertilitynew2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.05563 -0.03527 -0.01930  0.04337  0.07358
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  2.514944   0.537993   4.675 0.000165 ***
## fem.enroll. -0.023328   0.005240  -4.452 0.000273 ***
## urb.popul.  -0.016131   0.008942  -1.804 0.087126 .
## HHDI.CPI     0.006201   0.001440   4.307 0.000380 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04463 on 19 degrees of freedom
## Multiple R-squared:  0.5513, Adjusted R-squared:  0.4804
## F-statistic: 7.781 on 3 and 19 DF,  p-value: 0.001375
```

```
#life expectancy is found to be insignificant
ggqqplot(lmback2$residuals)
```




```
shapiro.test(lmback2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  lmback2$residuals
## W = 0.88631, p-value = 0.01334
```

```
#p=0.01334<0.05, the normality of residuals is not significant
```

```r
#lasso based on the full data
lambda_seq <- 10^seq(3, -3, by = -.1) # lambda sequence
lmlasso <- glmnet(Fertilityxnew2, Fertilitynew2$fertility, alpha = 1, lambda  = lambda_seq)
# Checking the model
summary(lmlasso)
```

```
##           Length Class     Mode
## a0         61    -none-    numeric
## beta      305    dgCMatrix S4
## df         61    -none-    numeric
## dim         2    -none-    numeric
## lambda     61    -none-    numeric
## dev.ratio  61    -none-    numeric
## nulldev     1    -none-    numeric
## npasses     1    -none-    numeric
## jerr        1    -none-    numeric
## offset      1    -none-    logical
## call        5    -none-    call
## nobs        1    -none-    numeric
```

```r
# Using cross validation glmnet
Fertilityxnew2 <- data.matrix(Fertilityxnew2)
lasso_cv <- cv.glmnet(Fertilityxnew2, Fertilitynew2$fertility, alpha = 1, lambda  = lambda_seq)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per
## fold
```

```r
# Best lambda value
best_lambda <- lasso_cv$lambda.min
best_lambda
```

```
## [1] 0.001
```

```r
# the best lambda 0.001995262
lmlassobest <- glmnet(Fertilityxnew2, Fertilitynew2$fertility, alpha = 1, lambda  = best_lambda)
summary(lmlassobest)
```

```
##           Length Class     Mode
## a0        1      -none-    numeric
## beta      5      dgCMatrix S4
## df        1      -none-    numeric
## dim       2      -none-    numeric
## lambda    1      -none-    numeric
## dev.ratio 1      -none-    numeric
## nulldev   1      -none-    numeric
## npasses   1      -none-    numeric
## jerr      1      -none-    numeric
## offset    1      -none-    logical
## call      5      -none-    call
## nobs      1      -none-    numeric
```

```
coef(lmlassobest)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                          s0
## (Intercept)    1.6336748758
## HHDI.CPI       0.0039977237
## fem.enroll.   -0.0180060069
## urb.popul.    -0.0058683465
## lifexpect      0.0028778104
## inf.mortality  0.0009928727
```

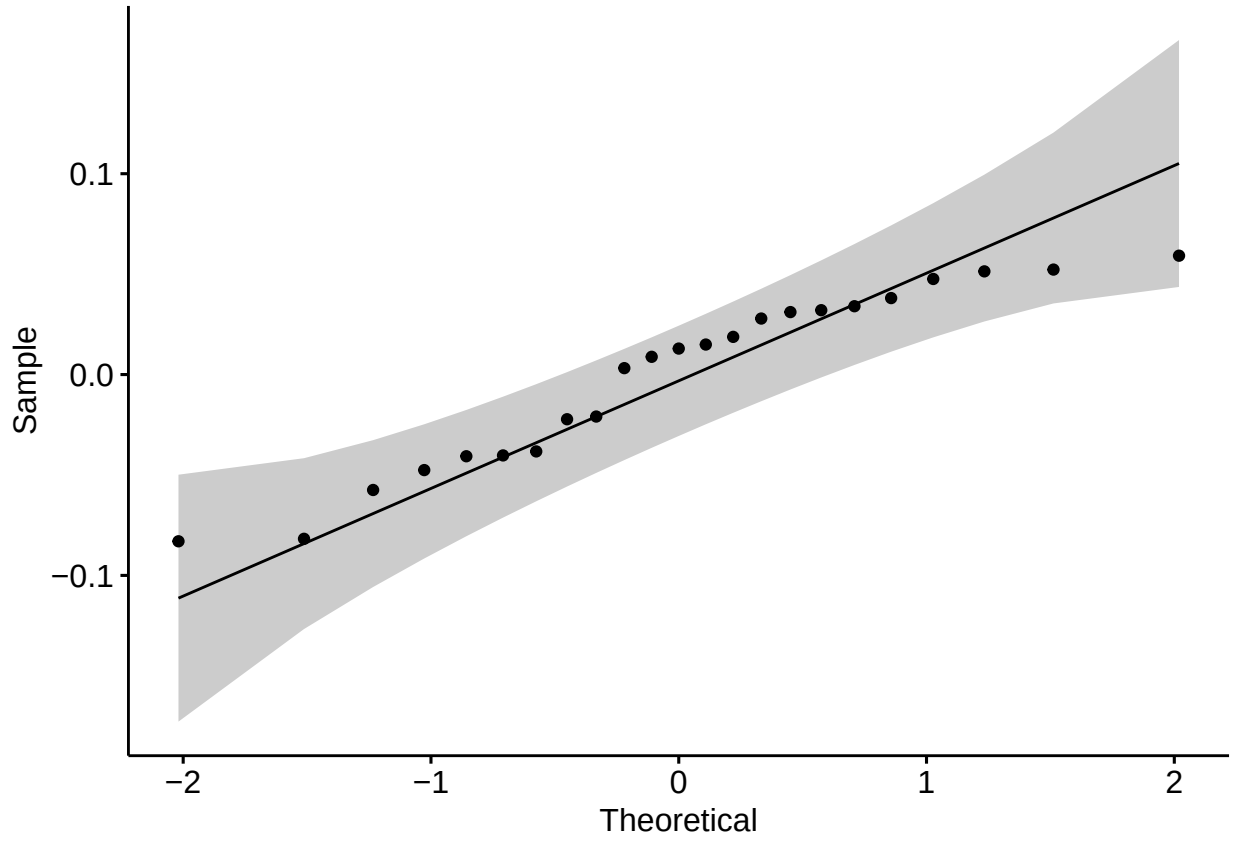
```
e <- predict(lmlassobest, Fertilityxnew2)-Fertilitynew2$fertility
e <- as.numeric(e)
class(e) # list to vector transformation
```

```
## [1] "numeric"
```

```
shapiro.test(e) #0.1759
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e
## W = 0.92537, p-value = 0.08687
```

```
ggqqplot(e)
```




```
# normality is acceptable
```

```
#ridge based on the full data
lambda_seq <- 10^seq(4, -4, by = -.001) # lambda sequence
lmridge <- glmnet(Fertilityxnew2, Fertilitynew2$fertility, alpha = 0, lambda  = lambda_seq)
# Checking the model
summary(lmridge)
```

```
##           Length Class     Mode
## a0         8001  -none-    numeric
## beta      40005  dgCMatrix S4
## df         8001  -none-    numeric
## dim           2  -none-    numeric
## lambda     8001  -none-    numeric
## dev.ratio  8001  -none-    numeric
## nulldev       1  -none-    numeric
## npasses       1  -none-    numeric
## jerr          1  -none-    numeric
## offset        1  -none-    logical
## call          5  -none-    call
## nobs          1  -none-    numeric
```

```
# Using cross validation glmnet
Fertilityxnew <- data.matrix(Fertilityxnew2)
ridge_cv <- cv.glmnet(Fertilityxnew2, Fertilitynew2$fertility, alpha = 0, lambda  = lambda_seq)
```

```
## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per
## fold
```

```
# Best lambda value
best_lambda2 <- ridge_cv$lambda.min
best_lambda2
```

```
## [1] 0.0001202264
```

```
# the best lambda 0.0001592209
lmridgebest <- glmnet(Fertilityxnew2, Fertilitynew2$fertility, alpha = 0, lambda  = best_lambda2)
summary(lmridgebest)
```

```
##           Length Class     Mode
## a0        1      -none-    numeric
## beta      5      dgCMatrix S4
## df        1      -none-    numeric
## dim       2      -none-    numeric
## lambda    1      -none-    numeric
## dev.ratio 1      -none-    numeric
## nulldev   1      -none-    numeric
## npasses   1      -none-    numeric
## jerr      1      -none-    numeric
## offset    1      -none-    logical
## call      5      -none-    call
## nobs      1      -none-    numeric
```

```
coef(lmridgebest)
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                            s0
## (Intercept)     0.018802959
## HHDI.CPI        0.005429286
## fem.enroll.    -0.024216609
## urb.popul.     -0.014196999
## lifexpect       0.030684595
## inf.mortality   0.010973568
```

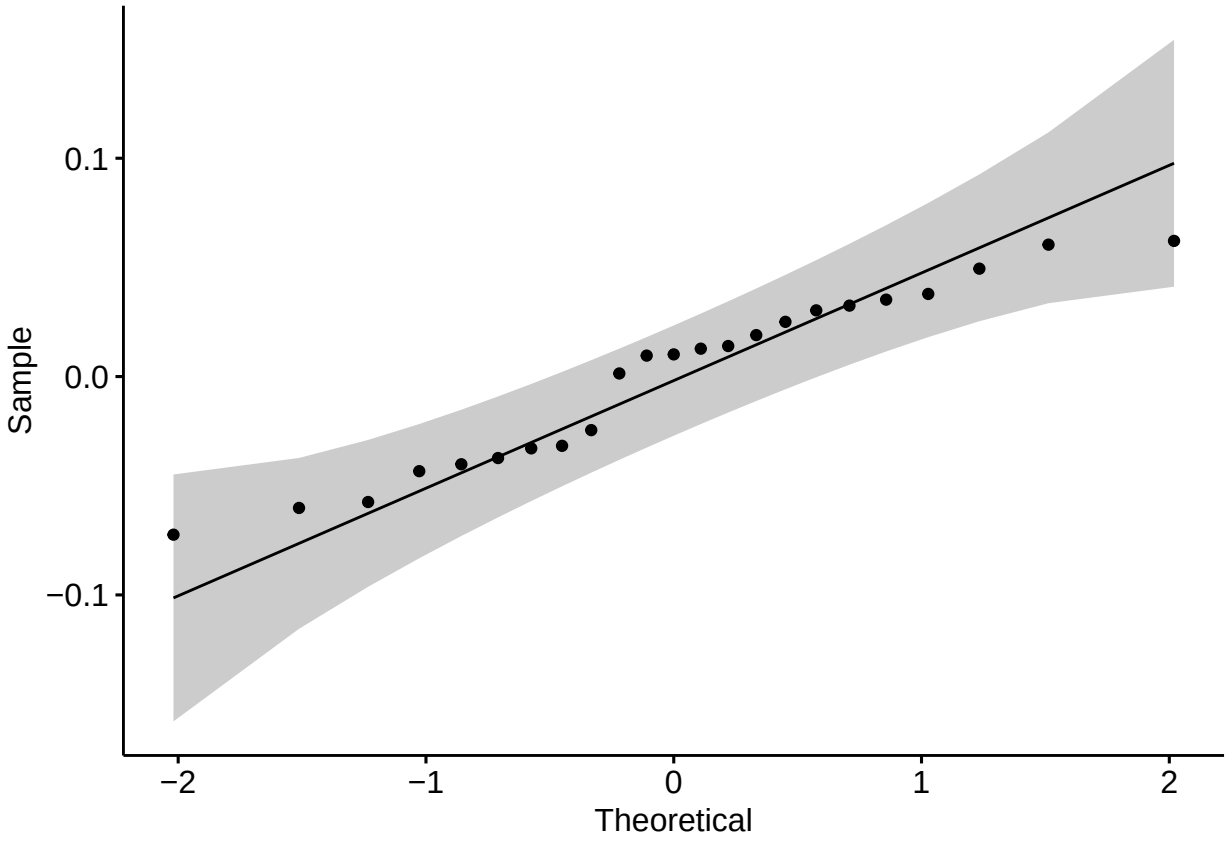
```
f <- predict(lmridgebest, Fertilityxnew2)-Fertilitynew2$fertility
f <- as.numeric(f)
class(f) # list to vector transformation
```

```
## [1] "numeric"
```

```
shapiro.test(f) #0.1979
```

```
##
##  Shapiro-Wilk normality test
##
## data:  f
## W = 0.94324, p-value = 0.2108
```

```
ggqqplot(f)
```




```
# normality is acceptable
```