Multiple Linear Regression(Swiss Data)

Suresh Kumar Prajapati

11-04-2024

```
rm(list=ls())
```

load the data

```
library(knitr)
```

```
## Warning: package 'knitr' was built under R version 4.3.3
```

```
data=swiss
head(data)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

```
kable(head(data))
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6

```
dim(swiss)
```

```
## [1] 47  6
```

to cheack the normality

```
head(swiss)
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
```

```
shapiro.test(data$Fertility)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$Fertility
## W = 0.97307, p-value = 0.3449
```

if p > alpha then we fail to reject the null hypothesis i.e data distributed normally. if p < alpha then we fail to accept the null hypothesis i.e data not distributed normally.

```
sample=sample(c(TRUE,FALSE),nrow(data),replace=TRUE,prob=c(0.7,0.3))
x=train=data[sample, ]
y=test=data[!sample, ]
kable(head(test))
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Franches-Mnt	92.5	39.7	5	5	93.40	20.2
Aigle	64.1	62.0	21	12	8.52	16.5
Nyone	56.6	50.9	22	12	15.14	16.7
Oron	72.5	71.2	12	1	2.40	21.0
Vevey	58.3	26.8	25	19	18.46	20.9
Herens	77.3	89.7	5	2	100.00	18.3

```
kable(head(train))
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
Courtelary	80.2	17.0	15	12	9.96	22.2
Delemont	83.1	45.1	6	9	84.84	22.2
Moutier	85.8	36.5	12	7	33.77	20.3
Neuveville	76.9	43.5	17	15	5.16	20.6
Porrentruy	76.1	35.3	9	7	90.57	26.6
Broye	83.8	70.2	16	7	92.85	23.6

```
str(x)
```

```
## 'data.frame':    37 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 85.8 76.9 76.1 83.8 92.4 82.4 82.9 87.1 ...
##  $ Agriculture     : num  17 45.1 36.5 43.5 35.3 70.2 67.8 53.3 45.2 64.5 ...
##  $ Examination     : int  15 6 12 17 9 16 14 12 16 14 ...
##  $ Education       : int  12 9 7 15 7 7 8 7 13 6 ...
##  $ Catholic        : num  9.96 84.84 33.77 5.16 90.57 ...
##  $ Infant.Mortality: num  22.2 22.2 20.3 20.6 26.6 23.6 24.9 21 24.4 24.5 ...
```

```
summary(train)
```

```
##    Fertility      Agriculture     Examination      Education
##  Min.   :42.80   Min.   : 7.70   Min.   : 3.00   Min.   : 2.00
##  1st Qu.:65.40   1st Qu.:35.30   1st Qu.:13.00   1st Qu.: 6.00
##  Median :70.50   Median :54.10   Median :16.00   Median : 8.00
##  Mean   :71.31   Mean   :49.52   Mean   :16.76   Mean   :10.03
##  3rd Qu.:79.30   3rd Qu.:64.90   3rd Qu.:22.00   3rd Qu.:12.00
##  Max.   :92.40   Max.   :85.90   Max.   :35.00   Max.   :32.00
##     Catholic     Infant.Mortality
##  Min.   : 2.15   Min.   :10.80
##  1st Qu.: 4.97   1st Qu.:19.10
##  Median :11.22   Median :20.20
##  Mean   :37.96   Mean   :20.36
##  3rd Qu.:91.38   3rd Qu.:22.40
##  Max.   :99.71   Max.   :26.60
```

```
library(knitr)
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.3.3
```

```
library(faraway)
```

```
## Warning: package 'faraway' was built under R version 4.3.3
```

```
## Warning in check_dep_version(): ABI version mismatch:
## lme4 was built with Matrix ABI version 1
## Current Matrix ABI version is 0
## Please re-install lme4 from source or restore original 'Matrix' package

##
## Attaching package: 'faraway'

## The following object is masked from 'package:psych':
##
##     logit
```

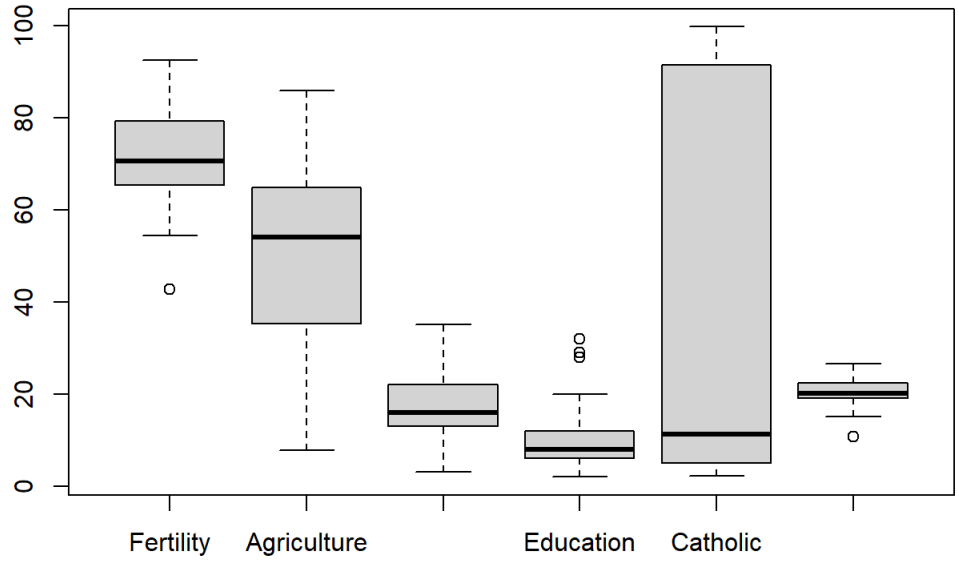
```
kable(round(describe(x)))
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Fertility	1	37	71	10	70	72	9	43	92	50	0	0	2
Agriculture	2	37	50	22	54	50	23	8	86	78	0	-1	4
Examination	3	37	17	7	16	16	6	3	35	32	0	0	1
Education	4	37	10	7	8	9	3	2	32	30	2	2	1
Catholic	5	37	38	42	11	35	12	2	100	98	1	-2	7
Infant.Mortality	6	37	20	3	20	21	3	11	27	16	-1	1	0

1. Initial data analysis that explores the numerical and graphical characteristics of data

Numerical characteristics

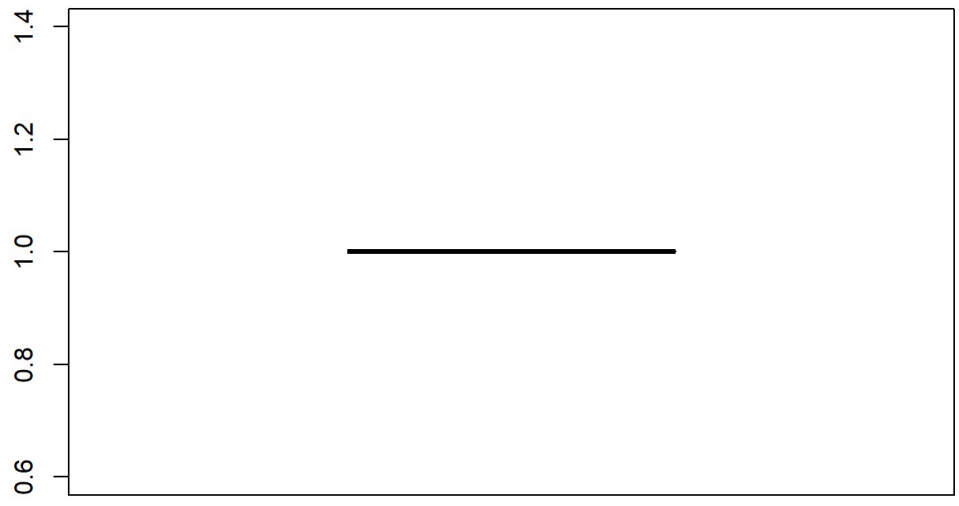
```
boxplot(train)
```




```
remove_out<-sample[!sample%in% boxplot.stats(sample)$out]
length(train)<-length(remove_out)
length(train)
```

```
## [1] 37
```

```
boxplot(remove_out)
```



Correlation analysis helps in selecting the most relevant features (variables) for the predictive modeling task.

The formula for Pearson's correlation coefficient is $r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$.

```
# Compute correlations between variables
correlation_matrix <- round(cor(x))
correlation_matrix
```

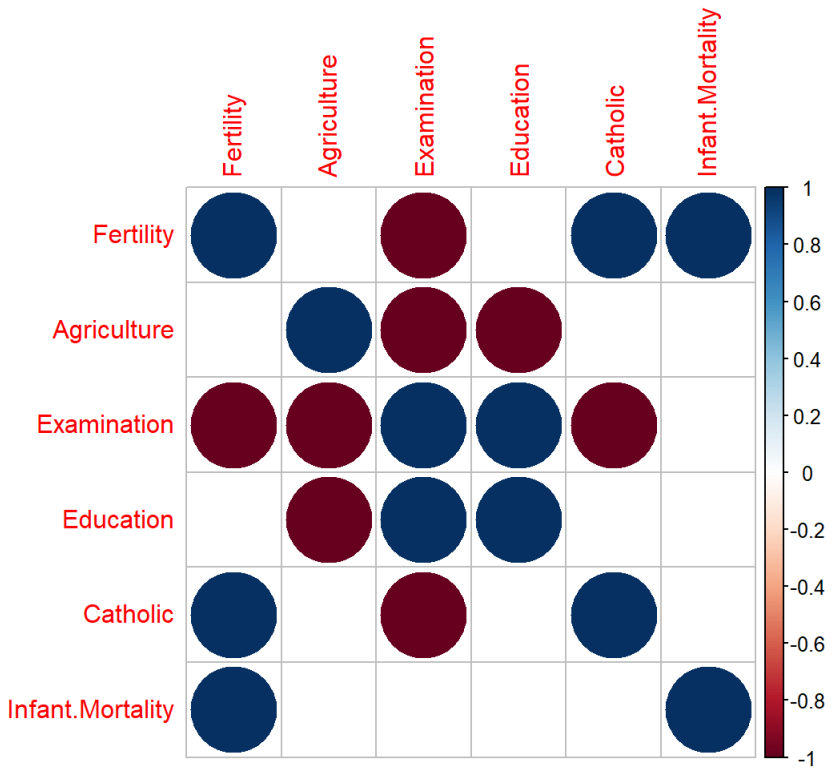
```
##                  Fertility Agriculture Examination Education Catholic
## Fertility                1           0          -1         0        1
## Agriculture              0           1          -1        -1        0
## Examination             -1          -1           1         1       -1
## Education                0          -1           1         1        0
## Catholic                 1           0          -1         0        1
## Infant.Mortality         1           0           0         0        0
##                  Infant.Mortality
## Fertility                       1
## Agriculture                     0
## Examination                     0
## Education                       0
## Catholic                        0
## Infant.Mortality                1
```

```
#install.packages("corrplot")
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

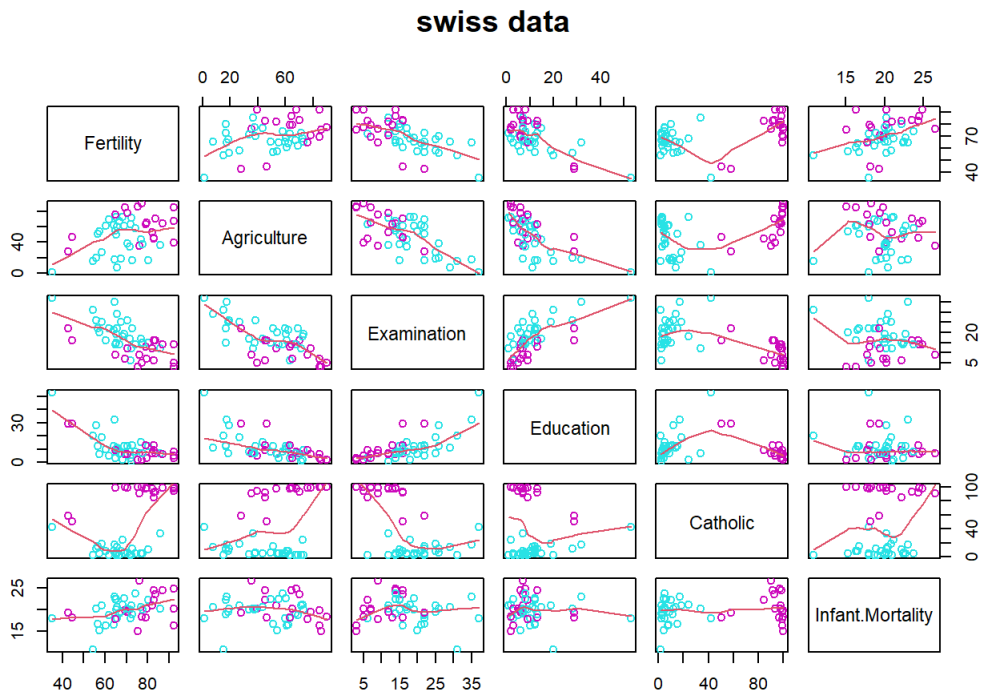
```
## corrplot 0.92 loaded
```

```
corrplot(correlation_matrix,method="circle")
```



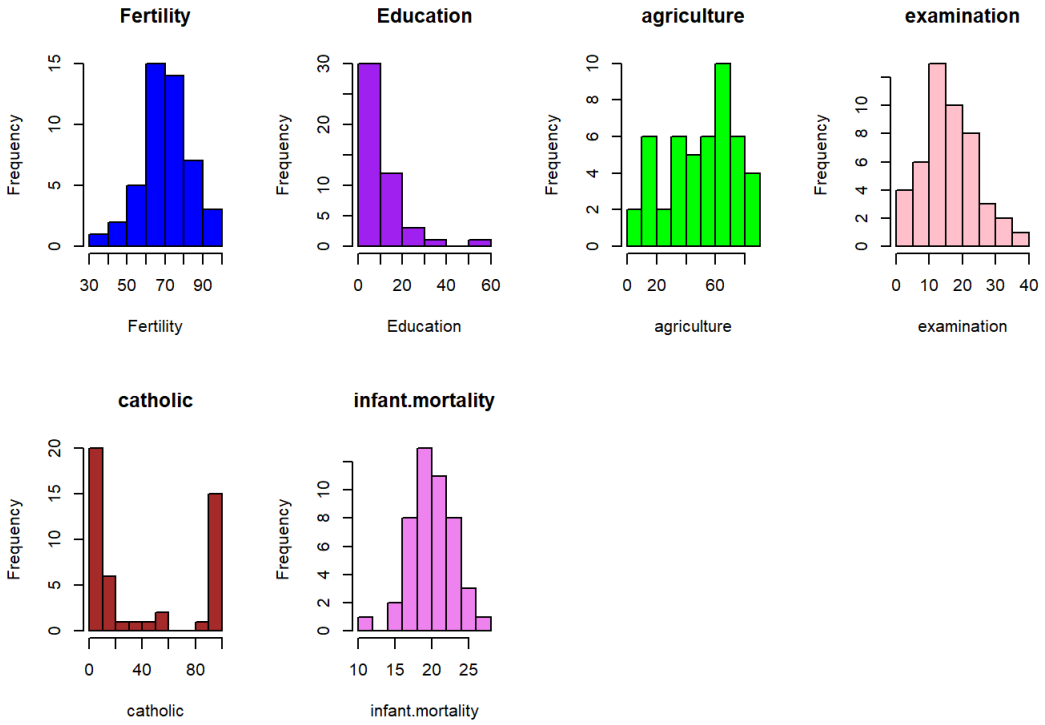
The numerical summary of the data shows that all the 6 variables are numerical with weak to moderate linear correlations among them.

```
pairs(swiss,panel=panel.smooth, main = "swiss data", col = 5 + (swiss$Catholic > 50))
```




```
par(mfrow=c(2,4))
hist(swiss$Fertility,main="Fertility",xlab="Fertility",col="blue")
hist(swiss$Education,main="Education",xlab="Education",col="purple")
hist(swiss$Agriculture,main="agriculture",xlab="agriculture",col="green")
```

```
hist(swiss$Examination,main="examination",xlab="examination",col="pink")
hist(swiss$Catholic,main="catholic",xlab="catholic",col="brown")

hist(swiss$Infant.Mortality,main="infant.mortality",xlab="infant.mortality",col="violet
")
```



nonlinearity issue # add

Polynomial terms(quadratic,cubic etc.)

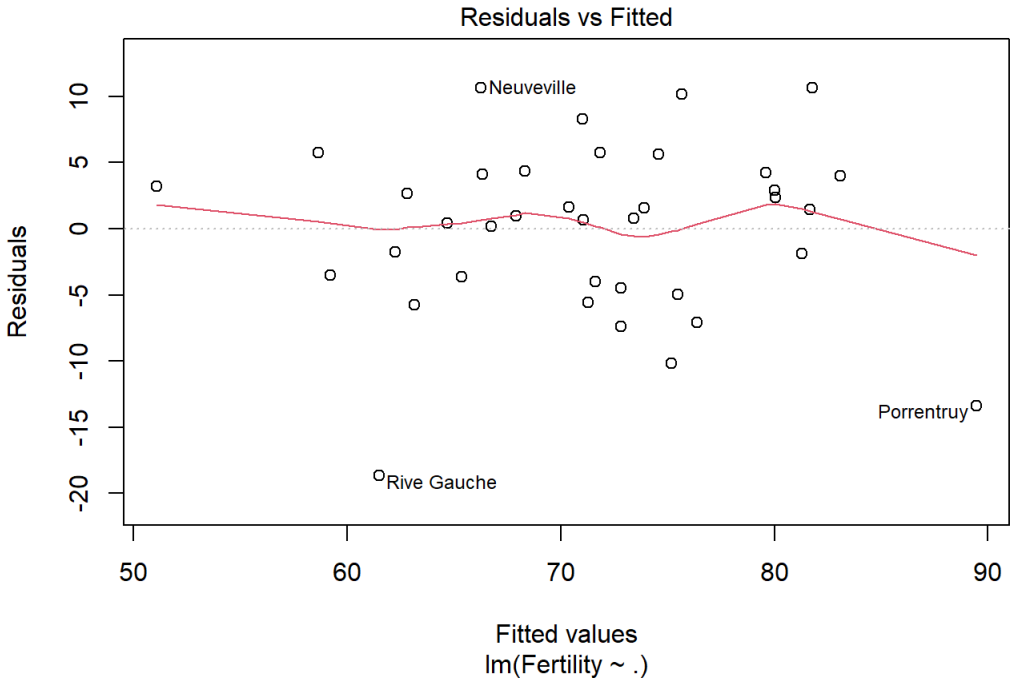
2. Variable selection to choose the best model

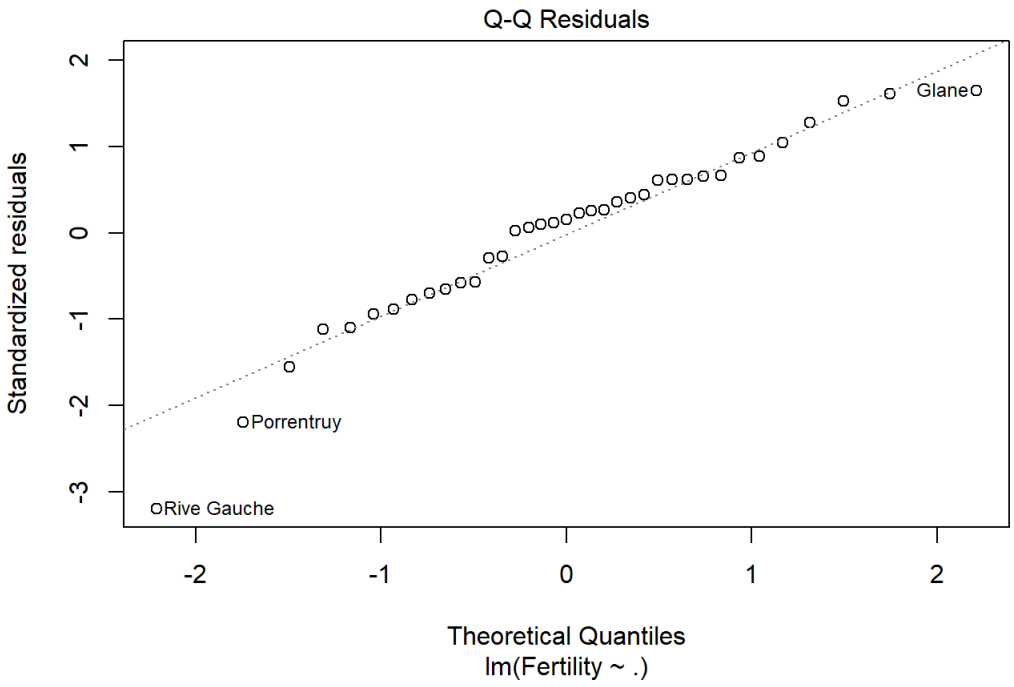
We start by fitting a linear regression model.

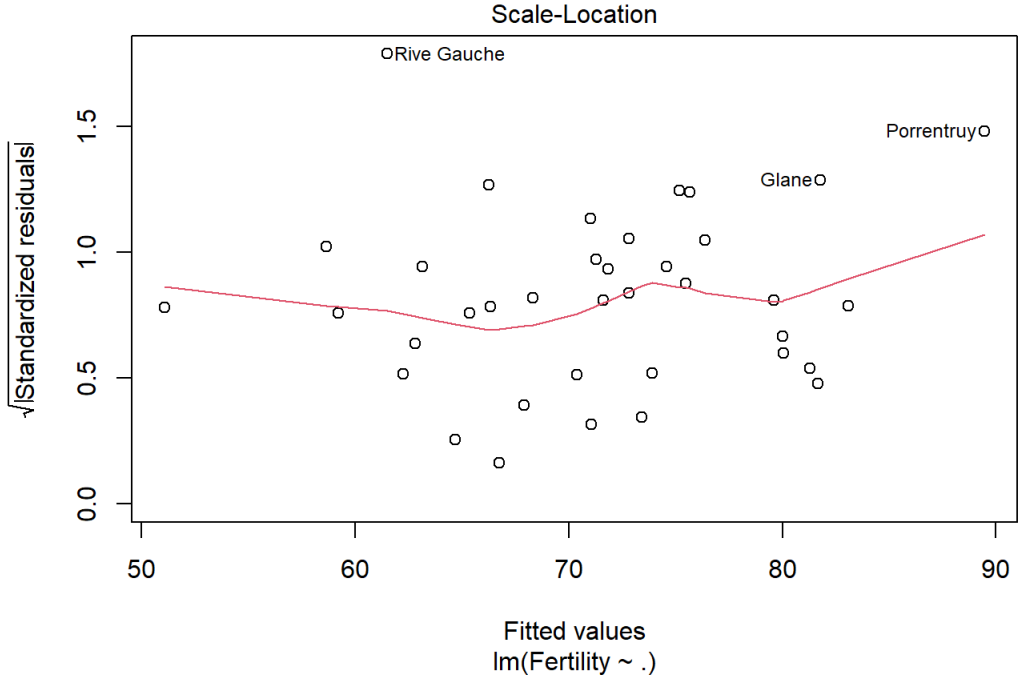
```
lmod <- lm(Fertility ~ ., x);
summary(lmod)

##
## Call:
## lm(formula = Fertility ~ ., data = x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.6794  -4.0258   0.9904   4.0828  10.6418
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      59.68625   11.26183   5.300 9.09e-06 ***
## Agriculture      -0.13329    0.07594  -1.755  0.08912 .
## Examination      -0.24284    0.27202  -0.893  0.37888
```

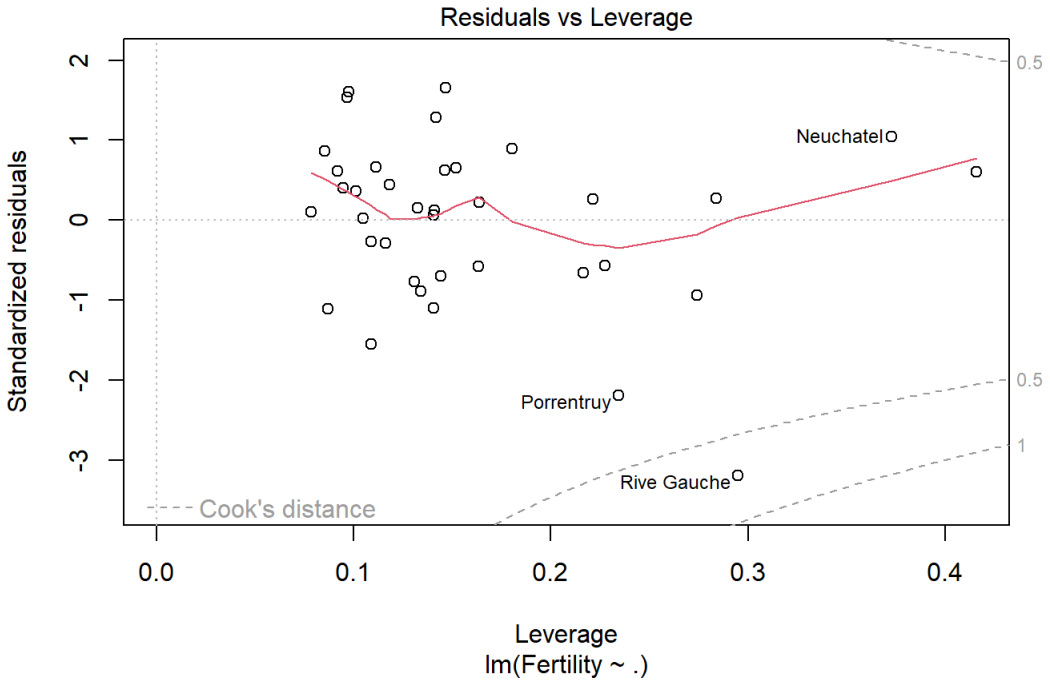
```
## Education        -0.63162    0.24890  -2.538  0.01641 *
## Catholic          0.08994    0.03670   2.451  0.02010 *
## Infant.Mortality  1.23827    0.40489   3.058  0.00456 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.957 on 31 degrees of freedom
## Multiple R-squared:  0.6068, Adjusted R-squared:  0.5434
## F-statistic: 9.568 on 5 and 31 DF,  p-value: 1.345e-05
```

```
plot(lmod)
```




Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(Fertility ~ .)
Glane
Porrentruy
Rive Gauche
-3
-2
-1
0
1
2


Scale-Location
√|Standardized residuals|
Fitted values
lm(Fertility ~ .)
Rive Gauche
Porrentruy
Glane
0.0
0.5
1.0
1.5
50
60
70
80
90




```
AIC_value=AIC(lmod)
AIC_value
```

```
## [1] 255.9921
```

```
vif(lmod)
```

```
##      Agriculture      Examination        Education         Catholic
##         1.985733         2.842057         2.277361         1.748733
## Infant.Mortality
##         1.124330
```

```
# Use drop1(lmod, test="F") alternatively
lmod_reduced = step(lmod)
```

```
## Start:  AIC=148.99
## Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     38.57 1538.8 147.93
## <none>                          1500.2 148.99
## - Agriculture       1    149.08 1649.3 150.50
## - Catholic          1    290.65 1790.9 153.54
## - Education         1    311.64 1811.9 153.97
## - Infant.Mortality  1    452.64 1952.9 156.75
##
```

```
## Step:  AIC=147.93
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## <none>                           1538.8 147.93
## - Agriculture       1    123.88 1662.7 148.79
## - Infant.Mortality  1    471.75 2010.5 155.82
## - Catholic          1    526.02 2064.8 156.81
## - Education         1    620.06 2158.8 158.46
```

```
summary(lmod_reduced)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.664  -5.452   1.511   3.214  11.781
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      55.01311    9.93966   5.535 4.19e-06 ***
## Agriculture      -0.11863    0.07391  -1.605  0.11830
## Education        -0.75112    0.20917  -3.591  0.00109 **
## Catholic          0.10582    0.03199   3.307  0.00233 **
## Infant.Mortality  1.26152    0.40277   3.132  0.00370 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.934 on 32 degrees of freedom
## Multiple R-squared:  0.5967, Adjusted R-squared:  0.5463
## F-statistic: 11.84 on 4 and 32 DF,  p-value: 5.169e-06
```

```
AIC_value=AIC(lmod_reduced)
AIC_value
```

```
## [1] 254.9313
```

```
vif(lmod_reduced)
```

```
##      Agriculture        Education         Catholic Infant.Mortality
##         1.892930         1.618670         1.337685         1.119679
```

```
anova(lmod,lmod_reduced)
```

```
## Analysis of Variance Table
##
## Model 1: Fertility ~ Agriculture + Examination + Education + Catholic +
##     Infant.Mortality
## Model 2: Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     31 1500.2
## 2     32 1538.8 -1   -38.569 0.797 0.3789
```

if p value less chosen significance level (0.05)i.e. there is a significant difference in the fit of the models.

if the p-value is greater than the significance level(0.05) i.e. there is no significant difference in the fit of the models.

By both a t-test and an ANOVA F test we find Examination does not have significant effect on Fertility.

We then treat Fertility ~ (Agriculture + Education + Catholic + Infant.Mortality)^2 as the full model, and use step() with BIC for selecting the best model. ## it is another way to fit the model

```
# Interaction term doesn't seem to bring major improvements
lmodi = lm(Fertility ~ (Agriculture + Education + Catholic +  Infant.Mortality)^2, data
= x)
lmodi
```

```
##
## Call:
## lm(formula = Fertility ~ (Agriculture + Education + Catholic +
##     Infant.Mortality)^2, data = x)
##
## Coefficients:
##                   (Intercept)                    Agriculture
##                     1.805e+02                     -1.753e+00
##                     Education                       Catholic
##                    -6.269e+00                      3.733e-01
##             Infant.Mortality         Agriculture:Education
##                    -4.626e+00                      3.038e-02
##         Agriculture:Catholic  Agriculture:Infant.Mortality
##                    -4.384e-04                      6.754e-02
##           Education:Catholic    Education:Infant.Mortality
##                    -1.463e-02                      2.439e-01
##    Catholic:Infant.Mortality
##                    -4.629e-03
```

```
lmodi_reduced = step(lmodi)
```

```
## Start:  AIC=142.38
## Fertility ~ (Agriculture + Education + Catholic + Infant.Mortality)^2
##
##                                Df Sum of Sq     RSS    AIC
## - Agriculture:Catholic          1     1.243  958.93 140.43
```

```
## - Catholic:Infant.Mortality     1     6.096  963.79 140.62
## <none>                                     957.69 142.38
## - Education:Catholic            1   152.094 1109.79 145.84
## - Agriculture:Infant.Mortality  1   162.724 1120.42 146.19
## - Education:Infant.Mortality    1   222.726 1180.42 148.12
## - Agriculture:Education         1   289.448 1247.14 150.16
##
## Step:  AIC=140.43
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality +
##     Agriculture:Education + Agriculture:Infant.Mortality + Education:Catholic +
##     Education:Infant.Mortality + Catholic:Infant.Mortality
##
##                                Df Sum of Sq     RSS    AIC
## - Catholic:Infant.Mortality     1     4.892  963.83 138.62
## <none>                                     958.93 140.43
## - Agriculture:Infant.Mortality  1   191.113 1150.05 145.16
## - Education:Catholic            1   195.428 1154.36 145.29
## - Education:Infant.Mortality    1   235.427 1194.36 146.56
## - Agriculture:Education         1   290.850 1249.78 148.23
##
## Step:  AIC=138.62
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality +
##     Agriculture:Education + Agriculture:Infant.Mortality + Education:Catholic +
##     Education:Infant.Mortality
##
##                                Df Sum of Sq     RSS    AIC
## <none>                                     963.83 138.62
## - Agriculture:Infant.Mortality  1    189.16 1152.98 143.25
## - Education:Catholic            1    196.35 1160.18 143.48
## - Education:Infant.Mortality    1    242.22 1206.04 144.91
## - Agriculture:Education         1    286.63 1250.46 146.25
```

```
summary(lmodi_reduced)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality + Agriculture:Education + Agriculture:Infant.Mortality +
##     Education:Catholic + Education:Infant.Mortality, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5894 -4.1033 -0.1738  2.3912  9.7976
##
## Coefficients:
```

```
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    177.892267  46.459959   3.829 0.000663 ***
## Agriculture                     -1.642570   0.556169  -2.953 0.006302 **
## Education                       -6.203632   1.980949  -3.132 0.004046 **
## Catholic                         0.243437   0.054023   4.506 0.000107 ***
## Infant.Mortality                -4.484545   2.205970  -2.033 0.051641 .
## Agriculture:Education            0.030114   0.010436   2.886 0.007439 **
## Agriculture:Infant.Mortality     0.061899   0.026406   2.344 0.026393 *
## Education:Catholic              -0.013940   0.005837  -2.388 0.023908 *
## Education:Infant.Mortality       0.240885   0.090808   2.653 0.013006 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.867 on 28 degrees of freedom
## Multiple R-squared:  0.7474, Adjusted R-squared:  0.6752
## F-statistic: 10.35 on 8 and 28 DF,  p-value: 1.31e-06
```

The fitted best model is Fertility = 53.75 − 0.134Agriculture − 0.515Education + 0.207Catholic + 1.24Infant.Mortality − 0.011Education:Catholic

with R2 = 0.7318 and Ra2 = 0.699.

```
#drop1(lmodi_reduced)
library(knitr)
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(car)

## Loading required package: carData

##
## Attaching package: 'car'

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

```
## The following objects are masked from 'package:faraway':
##
##     logit, vif

## The following object is masked from 'package:psych':
##
##     logit
```

```
summary(lmodi_reduced)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality + Agriculture:Education + Agriculture:Infant.Mortality +
##     Education:Catholic + Education:Infant.Mortality, data = x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5894 -4.1033 -0.1738  2.3912  9.7976
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  177.892267  46.459959   3.829 0.000663 ***
## Agriculture                   -1.642570   0.556169  -2.953 0.006302 **
## Education                     -6.203632   1.980949  -3.132 0.004046 **
## Catholic                       0.243437   0.054023   4.506 0.000107 ***
## Infant.Mortality              -4.484545   2.205970  -2.033 0.051641 .
## Agriculture:Education          0.030114   0.010436   2.886 0.007439 **
## Agriculture:Infant.Mortality   0.061899   0.026406   2.344 0.026393 *
## Education:Catholic            -0.013940   0.005837  -2.388 0.023908 *
## Education:Infant.Mortality     0.240885   0.090808   2.653 0.013006 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.867 on 28 degrees of freedom
## Multiple R-squared:  0.7474, Adjusted R-squared:  0.6752
## F-statistic: 10.35 on 8 and 28 DF,  p-value: 1.31e-06
```

```
prediction_fertility=predict(lmodi_reduced,newdata=y)
y=y%>% mutate(prediction_fertility=as.factor((prediction_fertility)))
kable(head(y))
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality	prediction_fertility
Franches-Mnt	92.5	39.7	5	5	93.40	20.2	87.249817897398
Aigle	64.1	62.0	21	12	8.52	16.5	61.6866110370566

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality	prediction_fertility
Nyone	56.6	50.9	22	12	15.14	16.7	65.3864866887445
Oron	72.5	71.2	12	1	2.40	21.0	60.8677189945037
Vevey	58.3	26.8	25	19	18.46	20.9	67.5404676865132
Herens	77.3	89.7	5	2	100.00	18.3	73.4623631738531

```
AIC_value=AIC(lmodi_reduced)
AIC_value
```

```
## [1] 245.6212
```

```
vif(lmodi_reduced)
```

```
## there are higher-order terms (interactions) in this model
## consider setting type = 'predictor'; see ?vif
```

```
##                  Agriculture                    Education
##                   149.737503                   202.803204
##                     Catholic             Infant.Mortality
##                     5.327834                    46.921780
##        Agriculture:Education Agriculture:Infant.Mortality
##                     4.424663                   145.616315
##           Education:Catholic   Education:Infant.Mortality
##                     6.004120                   183.676281
```

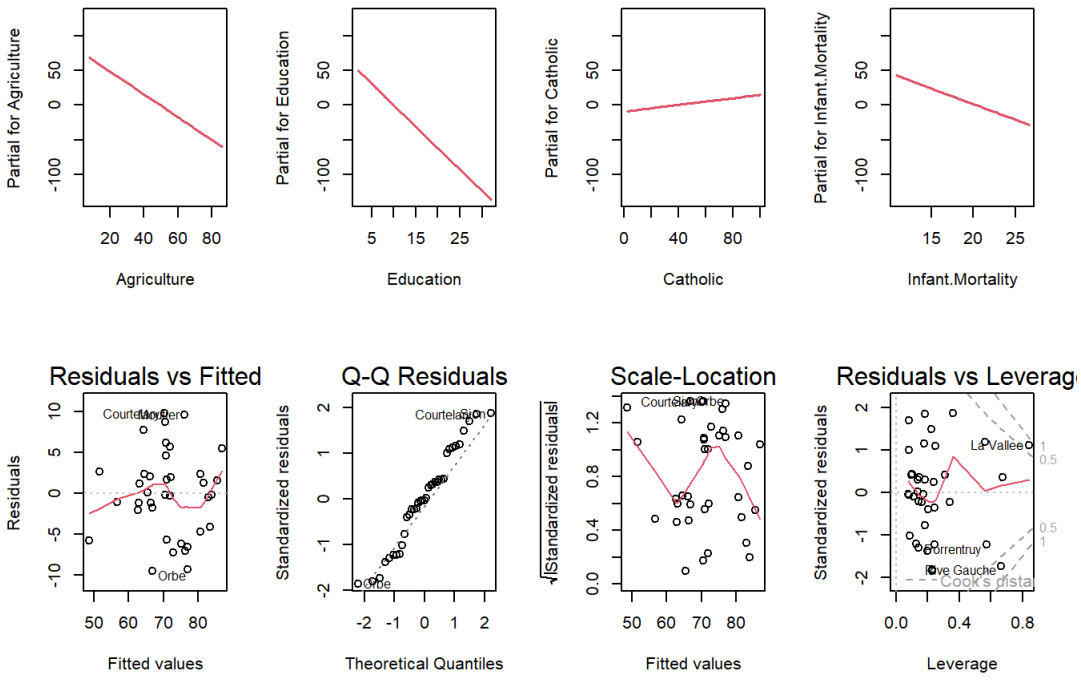
##The main purpose of the Normal Q-Q plot is to visually assess whether the distribution of the residuals from your model follows a normal (bell-shaped) distribution.

Interpretation Simplified: Straight Line: Ideal Scenario: If the points on the Q-Q plot fall approximately along the straight line, it suggests that the residuals are normally distributed. What It Means: This means that the residuals have a consistent spread around the mean, and most of the data points fall within a certain range. Deviation from the Line: Curved or Bent Line: If the points deviate from the straight line, it indicates that the residuals are not normally distributed. What It Means: This could suggest that the residuals have outliers, skewness, or heavy tails compared to a normal distribution.

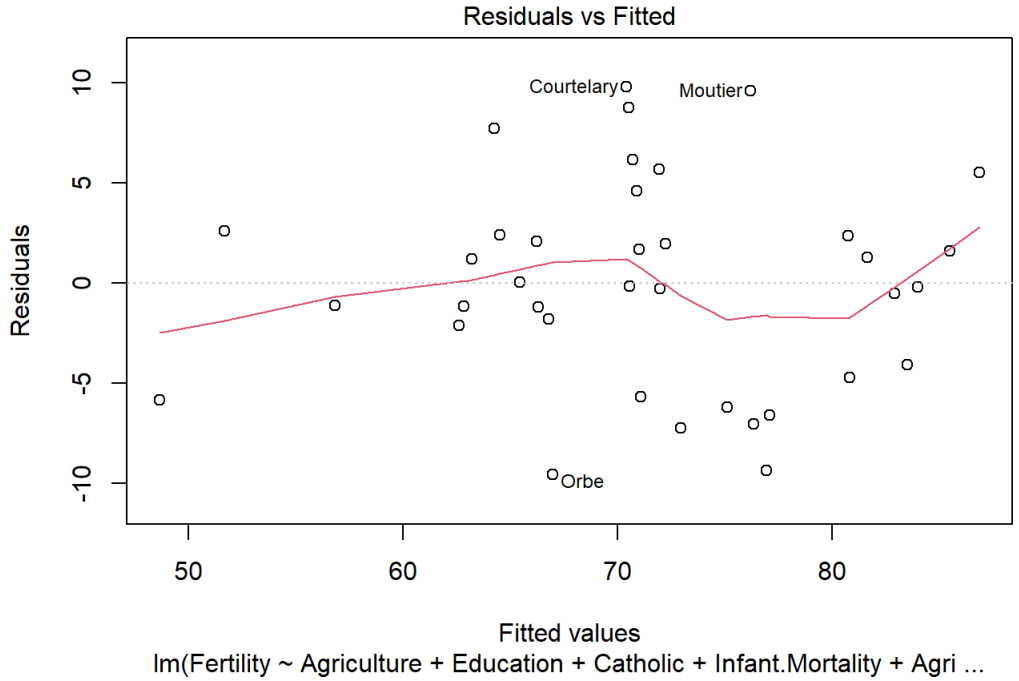
```
par(mfrow=c(2,4)); termplot(lmodi_reduced,partial=F,terms = NULL)
```

```
## Warning in termplot(lmodi_reduced, partial = F, terms = NULL): 'model' appears
## to involve interactions: see the help page
```

```
plot(lmodi_reduced)
```




```
# Create the Residuals vs Fitted plot
plot(lmodi_reduced, which = 1)
```



nonlinearity issue # add

Polynomial terms(quadratic,cubic etc.)

```
Catholic_sq <- train$Catholic^2
Catholic_sq
```

```
##  [1]   99.2016 7197.8256 1140.4129   26.6256 8202.9249 8621.1225 9440.0656
```

```
##  [8] 9539.4289 8350.3044 9723.9321    5.1529   19.6249    7.9524  585.6400
## [15]   10.8900  146.6521    4.6225    8.0656   27.3529   20.4304   17.6400
## [22]   27.3529    6.5536   59.5984   37.2100 9942.0841 9936.1024 9793.0816
## [29] 9647.1684 9376.0489   31.5844  190.1641  125.8884  286.2864   24.7009
## [36]   74.8225 3402.3889
```

```
# Check lengths of variables
length(train$Education)
```

```
## [1] 37
```

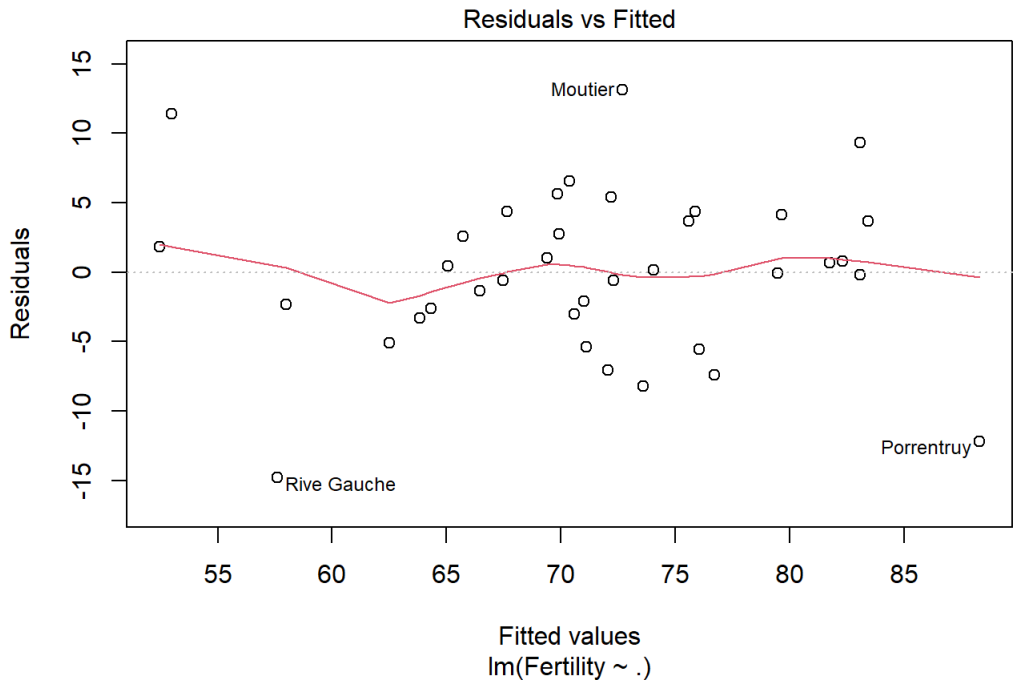
```
length(train$Education_sq)
```

```
## [1] 0
```

```
 #Add quadratic terms for all predictors
# Update the model
library(dplyr)
library(car)
library(psych)
library(faraway)
library(knitr)
Agriculture_sq <- train$Agriculture^2

Education_sq <- train$Education^2

Catholic_sq <- train$Catholic^2
Infant.Mortality_sq<-train$Infant.Mortality^2

library(dplyr)

z= mutate(x, Infant.Mortality_sq,Catholic_sq,Education_sq,Agriculture_sq )
head(z)
```

```
##            Fertility Agriculture Examination Education Catholic
## Courtelary      80.2        17.0          15        12     9.96
## Delemont        83.1        45.1           6         9    84.84
## Moutier         85.8        36.5          12         7    33.77
## Neuveville      76.9        43.5          17        15     5.16
## Porrentruy      76.1        35.3           9         7    90.57
## Broye           83.8        70.2          16         7    92.85
##            Infant.Mortality Infant.Mortality_sq Catholic_sq Education_sq
## Courtelary             22.2              492.84     99.2016          144
## Delemont               22.2              492.84   7197.8256           81
## Moutier                20.3              412.09   1140.4129           49
## Neuveville             20.6              424.36     26.6256          225
```

```
## Porrentruy             26.6              707.56   8202.9249           49
## Broye                  23.6              556.96   8621.1225           49
##            Agriculture_sq
## Courtelary         289.00
## Delemont          2034.01
## Moutier           1332.25
## Neuveville        1892.25
## Porrentruy        1246.09
## Broye             4928.04

lmodi_reduced_quad <- lm(Fertility ~ ., data = z);
lmodi_reduced_quad

##
## Call:
## lm(formula = Fertility ~ ., data = z)
##
## Coefficients:
##         (Intercept)          Agriculture          Examination
##           40.8618813           -0.0391517           -0.2421793
##           Education             Catholic     Infant.Mortality
##           0.5828343           -0.1019864            2.3620667
## Infant.Mortality_sq         Catholic_sq         Education_sq
##          -0.0286445            0.0020001           -0.0342583
##      Agriculture_sq
##          -0.0009223

plot(lmodi_reduced_quad,which=1)
```



$\text{AIC} = 2k - 2\ln(\hat{L})$

```
#prediction_fertility=predict(lmodi_reduced_quad,y)
#prediction_fertility
#y=y%>% mutate(prediction_fertility=as.factor((prediction_fertility)))
#kable(head(y)

AIC_value=AIC(lmodi_reduced_quad)
AIC_value

## [1] 257.5412

vif(lmodi_reduced_quad)

##         Agriculture         Examination           Education            Catholic
##           29.190337             3.201734            20.732133           92.522479
##    Infant.Mortality Infant.Mortality_sq        Catholic_sq        Education_sq
##           69.432928            70.534909            97.231819           18.002574
##      Agriculture_sq
##           29.896215
```

Apply Box-Cox Transformation can help stabilize the variance and make the relationship between the predictors and the response more linear

`

```
library(MASS)
```

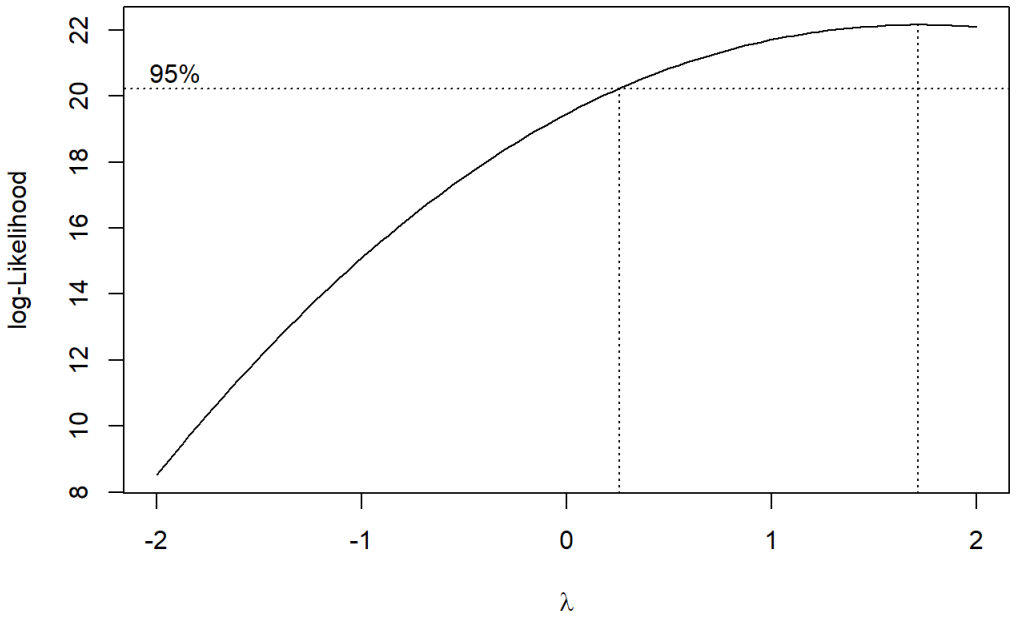
```
##
## Attaching package: 'MASS'
```

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

```
# Apply Box-Cox transformation to the response variable
lmodi_boxcox <- lm(Fertility ~ Agriculture + Education + Catholic + Infant.Mortality,
x)
 lmodi_boxcox
```

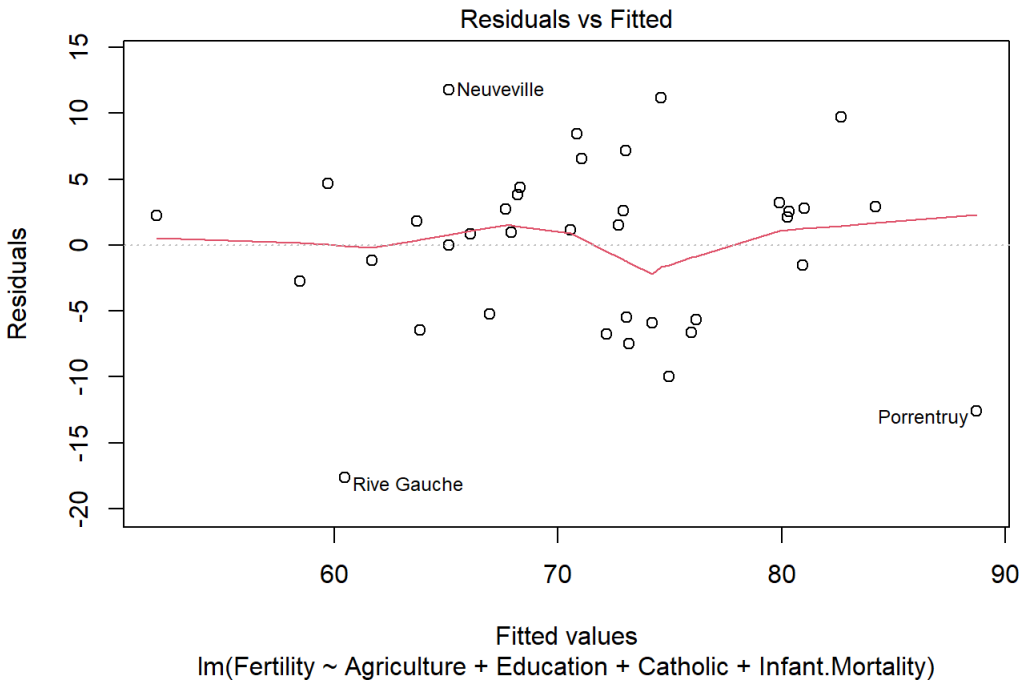
```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = x)
##
## Coefficients:
##      (Intercept)       Agriculture         Education          Catholic
##          55.0131           -0.1186           -0.7511            0.1058
## Infant.Mortality
##           1.2615
```

```
lambda <- boxcox(lmodi_boxcox)$lambda
```




```
lambda
```

```
## NULL

# Update the model with the transformed response variable
lmodi_reduced_boxcox <- lm(Fertility ~ Agriculture + Education + Catholic +
Infant.Mortality, x)
lmodi_reduced_boxcox

##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = x)
##
## Coefficients:
##      (Intercept)       Agriculture         Education          Catholic
##          55.0131           -0.1186           -0.7511            0.1058
## Infant.Mortality
##           1.2615

plot(lmodi_reduced_boxcox,which=1)
```




```
summary(lmodi_reduced_boxcox)

##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = x)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.664  -5.452   1.511   3.214  11.781
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      55.01311    9.93966   5.535 4.19e-06 ***
## Agriculture      -0.11863    0.07391  -1.605  0.11830
## Education        -0.75112    0.20917  -3.591  0.00109 **
## Catholic          0.10582    0.03199   3.307  0.00233 **
## Infant.Mortality  1.26152    0.40277   3.132  0.00370 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.934 on 32 degrees of freedom
## Multiple R-squared:  0.5967, Adjusted R-squared:  0.5463
## F-statistic: 11.84 on 4 and 32 DF,  p-value: 5.169e-06
```

```
prediction_fertility=predict(lmodi_reduced_boxcox,newdata=y)
y=y%>% mutate(prediction_fertility=as.factor((prediction_fertility)))
kable(head(y))
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality	prediction_fertility
Franches-Mnt	92.5	39.7	5	5	93.40	20.2	87.249817897398
Aigle	64.1	62.0	21	12	8.52	16.5	61.6866110370566
Nyone	56.6	50.9	22	12	15.14	16.7	65.3864866887445
Oron	72.5	71.2	12	1	2.40	21.0	60.8677189945037
Vevey	58.3	26.8	25	19	18.46	20.9	67.5404676865132
Herens	77.3	89.7	5	2	100.00	18.3	73.4623631738531

```
AIC_value=AIC(lmodi_reduced_boxcox)
AIC_value
```

```
## [1] 254.9313
```

```
vif(lmodi_reduced_boxcox)
```

```
##      Agriculture        Education         Catholic Infant.Mortality
##         1.892930         1.618670         1.337685         1.119679
```

For cheacking the hetroscedasticity issue

```
# Fit WLS model
weights <- 1 / residuals(lmodi_reduced)^2
```

```
weights
```

```
##   Courtelary     Delemont      Moutier   Neuveville   Porrentruy        Broye
##   0.01041749   0.18243942   0.01082947   0.02629059   0.04485689  23.08580012
##        Glane      Gruyere       Sarine      Veveyse      Aubonne     Avenches
##   0.03266093   3.73009921   0.62431970   0.38655545   0.17489653   0.02579931
##     Cossonay    Echallens     Grandson     Lausanne    La Vallee       Lavaux
##   0.76090390   0.23457263  12.35247131   0.80047990   0.14687995   0.67884658
##       Morges       Moudon         Orbe      Payerne Paysd'enhaut        Rolle
## 381.06438637   0.30955160   0.01087466   0.26098462   0.01672741   0.22432656
##      Yverdon      Conthey    Entremont     Martigwy      Monthey         Sion
##   0.03095828   0.04730509   0.02005029   0.02288976   0.05939336   0.01300647
##       Boudry La Chauxdfnd     Le Locle    Neuchatel   Val de Ruz ValdeTravers
##  33.10956333   0.01900016   0.35211239   0.69652239   0.03119922   0.01141878
##  Rive Gauche
##   0.02911413
```

```
lmodi_wls=model.frame.default(formula = Fertility ~ Agriculture + Education +
Catholic + Infant.Mortality, data = x, weights = weights,
drop.unused.levels = TRUE)
head(lmodi_wls)
```

```
##            Fertility Agriculture Education Catholic Infant.Mortality
## Courtelary      80.2        17.0        12     9.96             22.2
## Delemont        83.1        45.1         9    84.84             22.2
## Moutier         85.8        36.5         7    33.77             20.3
## Neuveville      76.9        43.5        15     5.16             20.6
## Porrentruy      76.1        35.3         7    90.57             26.6
## Broye           83.8        70.2         7    92.85             23.6
##              (weights)
## Courtelary  0.01041749
## Delemont    0.18243942
## Moutier     0.01082947
## Neuveville  0.02629059
## Porrentruy  0.04485689
## Broye      23.08580012
```

```
lmodi_wl=lm(Fertility ~ .,x)
lmodi_wl
```

```
##
## Call:
## lm(formula = Fertility ~ ., data = x)
##
## Coefficients:
##      (Intercept)       Agriculture       Examination         Education
```

```
##          59.68625           -0.13329           -0.24284           -0.63162
##          Catholic  Infant.Mortality
##           0.08994            1.23827
```

```
prediction_fertility=predict(lmodi_wl,newdata=y)
y=y%>% mutate(prediction_fertility=as.factor((prediction_fertility)))
kable(head(y))
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality	prediction_fertility
Franches-Mnt	92.5	39.7	5	5	93.40	20.2	87.249817897398
Aigle	64.1	62.0	21	12	8.52	16.5	61.6866110370566
Nyone	56.6	50.9	22	12	15.14	16.7	65.3864866887445
Oron	72.5	71.2	12	1	2.40	21.0	60.8677189945037
Vevey	58.3	26.8	25	19	18.46	20.9	67.5404676865132
Herens	77.3	89.7	5	2	100.00	18.3	73.4623631738531

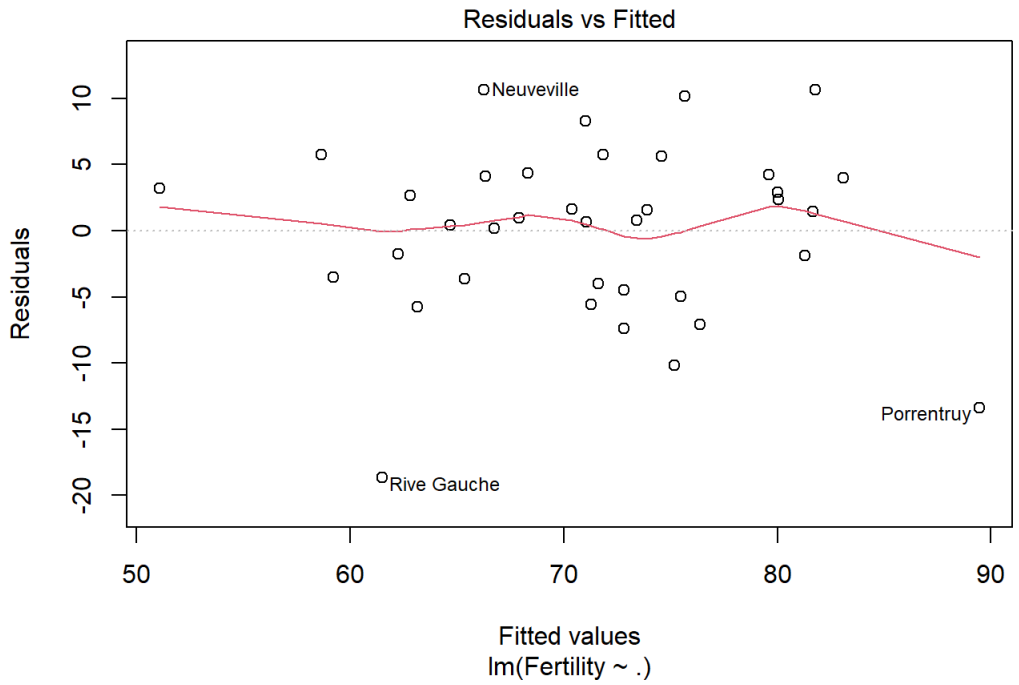
```
AIC_value=AIC(lmodi_wl)
AIC_value
```

```
## [1] 255.9921
```

```
vif(lmodi_wl)
```

```
##      Agriculture      Examination        Education         Catholic
##         1.985733         2.842057         2.277361         1.748733
## Infant.Mortality
##         1.124330
```

```
plot(lmodi_wl, which=1)
```



Transform the Response

Variance for using logarithmic or square root transformation can help stabilize the variance

```
# Log transformation of the response variable
x$Fertility_log <- log(x$Fertility)
x$Fertility_log
```

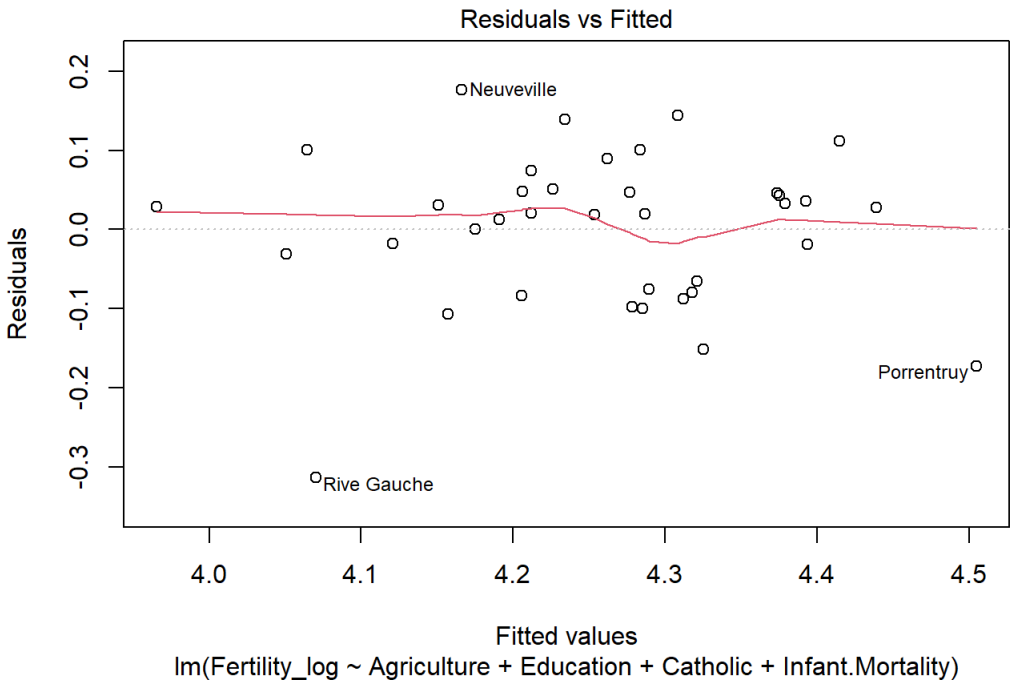
```
##  [1] 4.384524 4.420045 4.452019 4.342506 4.332048 4.428433 4.526127 4.411585
##  [9] 4.417635 4.467057 4.203199 4.232656 4.122284 4.223910 4.272491 4.019980
## [17] 3.994524 4.175925 4.182050 4.174387 4.050044 4.306764 4.276666 4.102643
## [25] 4.180522 4.324133 4.238445 4.255613 4.374498 4.373238 4.254193 4.185099
## [33] 4.286341 4.165114 4.351567 4.213608 3.756538
```

```
# Fit the model with the transformed response variable
lmodi_reduced_log <- lm(Fertility_log ~ Agriculture + Education + Catholic +
Infant.Mortality, x)
summary(lmodi_reduced_log)
```

```
##
## Call:
## lm(formula = Fertility_log ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31358 -0.07563  0.02055  0.04792  0.17623
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)       4.028429   0.149738  26.903  < 2e-16 ***
## Agriculture      -0.001685   0.001113  -1.513 0.140054
## Education        -0.012139   0.003151  -3.852 0.000529 ***
## Catholic          0.001346   0.000482   2.792 0.008764 **
## Infant.Mortality  0.018751   0.006067   3.090 0.004118 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1045 on 32 degrees of freedom
## Multiple R-squared:  0.5842, Adjusted R-squared:  0.5322
## F-statistic: 11.24 on 4 and 32 DF,  p-value: 8.27e-06
```

```
plot(lmodi_reduced_log,which=1)
```




```
AIC_value=AIC(lmodi_reduced_log)
AIC_value
```

```
## [1] -55.52851
```

```
vif(lmodi_reduced_log)
```

```
##      Agriculture        Education         Catholic Infant.Mortality
##         1.892930         1.618670         1.337685         1.119679
```

```
library(dplyr)
prediction_fertility=predict(lmodi_reduced_log,newdata=y)
y=y%>% mutate(prediction_fertility=as.factor((prediction_fertility)))
```

```
kable(head(y))
```

	Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality	prediction_fertility
Franches-Mnt	92.5	39.7	5	5	93.40	20.2	87.249817897398
Aigle	64.1	62.0	21	12	8.52	16.5	61.6866110370566
Nyone	56.6	50.9	22	12	15.14	16.7	65.3864866887445
Oron	72.5	71.2	12	1	2.40	21.0	60.8677189945037
Vevey	58.3	26.8	25	19	18.46	20.9	67.5404676865132
Herens	77.3	89.7	5	2	100.00	18.3	73.4623631738531

Conclusion

1:-VIF (Variance inflation Factor) for check the multicollinearity $VIF=\frac{1}{1-R^2}$ if VIF value less than 10 then less multicollinearity & if greater than 10 then multicollinearity occurrence.

AIC (Akaike Information Criterion):- measure for goodness of fit if the value of AIC less then no complexity occurrence then model performance is better.

$ AIC={-2()+2$

Here, model1 =lmod

```
                    AIC=241.3941
      vif(lmod)
 Agriculture      Examination        Education         Catholic
     1.778175         3.166950         2.586892         1.671764
Infant.Mortality
   1.191916


model2=lmod_reduced
                    AIC=276.15
  vif(lmod_reduced)

 Agriculture        Education         Catholic Infant.Mortality
    2.224731         1.871269         1.325160         1.092146



 model3=lmodi_reduced
                       AIC=273.5431
 vif(lmodi_reduced)
      Agriculture          Education           Catholic   Infant.Mortality
Education:Catholic
         2.294928           7.057182           3.416303           1.162189
```

7.143829

model4=lmodi_reduced_quad

AIC = 280.6484

Agriculture	Examination	Education	Catholic	Infant.Mortality
31.488752	4.153044	17.180881	85.814544	70.820632

Infant.Mortality_sq	Catholic_sq	Education_sq	Agriculture_sq
71.618003	89.931591	12.955854	31.087917

model5=lmodi_reduced_boxcox

AIC= 276.15

VIF=

Agriculture	Education	Catholic	Infant.Mortality
2.224731	1.871269	1.325160	1.092146

model6= lmodi_wls

AIC= 277.2222

VIF=

Agriculture	Examination	Education	Catholic	Infant.Mortality
2.402952	3.556452	2.811778	1.834457	1.093074

model7=lmodi_reduced_log

AIC= -52.45299

VIF= Agriculture Education Catholic Infant.Mortality 2.224731 1.871269 1.325160 1.092146

therefore model1,model2,model3,model5,model6, and model7 are good as compare to other.