```
> # Step 2: Load data, prepare the data for model fitting
>
> setwd('C:/users/mg/Desktop/Data Analytics/HW/PROJECT/project')
> getwd()
[1] "C:/users/mg/Desktop/Data Analytics/HW/PROJECT/project"
> mydata = read.table("Life Expectancy Data.csv", header=T, sep=",")
> set.seed(1234)
> head(mydata)
      Country Year     Status Life.expectancy Adult.Mortality infant.deaths Alcohol percentage.expenditure Hepatitis.B Measles  BMI under.five.deaths Polio
1 Afghanistan 2015 Developing            65.0             263           62    0.01              71.279624          65    1154 19.1                83     6
2 Afghanistan 2014 Developing            59.9             271           64    0.01              73.523582          62     492 18.6                86    58
3 Afghanistan 2013 Developing            59.9             268           66    0.01              73.219243          64     430 18.1                89    62
4 Afghanistan 2012 Developing            59.5             272           69    0.01              78.184215          67    2787 17.6                93    67
5 Afghanistan 2011 Developing            59.2             275           71    0.01               7.097109          68    3013 17.2                97    68
6 Afghanistan 2010 Developing            58.8             279           74    0.01              79.679367          66    1989 16.7               102    66
  Total.expenditure Diphtheria HIV.AIDS      GDP Population thinness..1.19.years thinness.5.9.years Income.composition.of.resources Schooling
1              8.16         65      0.1 584.25921   33736494                17.2                17.3                           0.479      10.1
2              8.18         62      0.1 612.69651     327582                17.5                17.5                           0.476      10.0
3              8.13         64      0.1 631.74498   31731688                17.7                17.7                           0.470       9.9
4              8.52         67      0.1 669.95900    3696958                17.9                18.0                           0.463       9.8
5              7.87         68      0.1  63.53723    2978599                18.2                18.2                           0.454       9.5
6              9.20         66      0.1 553.32894    2883167                18.4                18.4                           0.448       9.2
```

```
> str(mydata)
'data.frame':   2914 obs. of  22 variables:
 $ Country                        : Factor w/ 214 levels "Afghanistan",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
 $ Status                         : Factor w/ 2 levels "Developed","Developing": 2 2 2 2 2 2 2 2 2 2 ...
 $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
 $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
 $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
 $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
 $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
 $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
 $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
 $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
 $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
 $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
 $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
 $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
 $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
 $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
 $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
 $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
 $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
 $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405 ...
 $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
```

```
> # Step 2-1: Deal with missing values
> summary(mydata) # factor which has missing values: lfey, atmy, achl, hb, bmi, plio, totalex, dpria, gdp, pplon, th119, th59, iccr, sch.
             Country          Year            Status     Life.expectancy Adult.Mortality infant.deaths      Alcohol       percentage.expenditure
 Afghanistan    :  16   Min.   :2000   Developed : 512   Min.   :36.30   Min.   :  1.0   Min.   :   1.00   Min.   : 0.010   Min.   :    0.10
 Albania        :  16   1st Qu.:2004   Developing:2402   1st Qu.:63.20   1st Qu.: 73.0   1st Qu.:   2.00   1st Qu.: 0.860   1st Qu.:   36.25
 Algeria        :  16   Median :2008                     Median :72.15   Median :143.0   Median :   9.00   Median : 3.790   Median :  155.19
 Angola         :  16   Mean   :2008                     Mean   :69.29   Mean   :163.9   Mean   :  42.78   Mean   : 4.613   Mean   :  932.09
 Antigua and Barbuda:  16   3rd Qu.:2012                  3rd Qu.:75.70   3rd Qu.:227.0   3rd Qu.:  33.00   3rd Qu.: 7.755   3rd Qu.:  630.22
 Argentina      :  16   Max.   :2015                     Max.   :89.00   Max.   :723.0   Max.   :1800.00   Max.   :17.870   Max.   :19479.91
 (Other)        :2818                                    NA's   :10      NA's   :10      NA's   :848       NA's   :191      NA's   :587
  Hepatitis.B       Measles             BMI        under.five.deaths      Polio       Total.expenditure   Diphtheria       HIV.AIDS            GDP
 Min.   : 2.00   Min.   :    1   Min.   : 1.00   Min.   :   1.00   Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100   Min.   :     1.68
 1st Qu.:77.00   1st Qu.:   17   1st Qu.:19.38   1st Qu.:   3.00   1st Qu.:78.00   1st Qu.: 4.260    1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   463.94
 Median :92.00   Median :  128   Median :43.80   Median :  10.00   Median :93.00   Median : 5.760    Median :93.00   Median : 0.100   Median :  1766.95
 Mean   :81.09   Mean   : 3658   Mean   :38.47   Mean   :  57.58   Mean   :82.61   Mean   : 5.945    Mean   :82.43   Mean   : 1.743   Mean   :  7483.16
 3rd Qu.:97.00   3rd Qu.: 1207   3rd Qu.:56.30   3rd Qu.:  46.00   3rd Qu.:97.00   3rd Qu.: 7.500    3rd Qu.:97.00   3rd Qu.: 0.800   3rd Qu.:  5910.81
 Max.   :99.00   Max.   :212183  Max.   :87.30   Max.   :2500.00   Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600   Max.   :119172.74
 NA's   :551     NA's   :976     NA's   :34      NA's   :785       NA's   :19      NA's   :216       NA's   :19                       NA's   :424
   Population       thinness..1.19.years thinness.5.9.years Income.composition.of.resources   Schooling
 Min.   :3.400e+01   Min.   : 0.100       Min.   : 0.10      Min.   :0.2530                  Min.   : 2.80
 1st Qu.:1.958e+05   1st Qu.: 1.600       1st Qu.: 1.50      1st Qu.:0.5250                  1st Qu.:10.20
 Median :1.387e+06   Median : 3.300       Median : 3.30      Median :0.6870                  Median :12.40
 Mean   :1.275e+07   Mean   : 4.829       Mean   : 4.86      Mean   :0.6589                  Mean   :12.12
 3rd Qu.:7.420e+06   3rd Qu.: 7.200       3rd Qu.: 7.20      3rd Qu.:0.7840                  3rd Qu.:14.30
 Max.   :1.294e+09   Max.   :27.700       Max.   :28.60      Max.   :0.9480                  Max.   :20.70
 NA's   :628         NA's   :34           NA's   :34         NA's   :281                     NA's   :175
```

```
> summary(newdata)
             Country          Year            Status     Life.expectancy Adult.Mortality infant.deaths      Alcohol       percentage.expenditure
 Afghanistan    :  16   Min.   :2000   Developed : 512   Min.   :36.30   Min.   :  1.0   Min.   :   1.00   Min.   : 0.010   Min.   :    0.1
 Albania        :  16   1st Qu.:2004   Developing:2376   1st Qu.:63.40   1st Qu.: 73.0   1st Qu.:   3.00   1st Qu.: 0.910   1st Qu.:   45.7
 Algeria        :  16   Median :2008                     Median :72.20   Median :142.0   Median :  23.00   Median : 3.795   Median :  242.7
 Angola         :  16   Mean   :2007                     Mean   :69.38   Mean   :162.9   Mean   :  41.21   Mean   : 4.623   Mean   :  927.1
 Antigua and Barbuda:  16   3rd Qu.:2011                  3rd Qu.:75.72   3rd Qu.:225.0   3rd Qu.:  41.21   3rd Qu.: 7.770   3rd Qu.:  927.1
 Argentina      :  16   Max.   :2015                     Max.   :89.00   Max.   :723.0   Max.   :1800.00   Max.   :17.870   Max.   :19479.9
 (Other)        :2792
  Hepatitis.B       Measles             BMI        under.five.deaths      Polio       Total.expenditure   Diphtheria       HIV.AIDS            GDP
 Min.   : 2.00   Min.   :    1   Min.   : 1.00   Min.   :   1.00   Min.   : 3.00   Min.   : 0.370    Min.   : 2.00   Min.   : 0.100   Min.   :     1.68
 1st Qu.:74.00   1st Qu.:   15   1st Qu.:19.40   1st Qu.:   4.00   1st Qu.:78.00   1st Qu.: 4.290    1st Qu.:78.00   1st Qu.: 0.100   1st Qu.:   550.50
 Median :88.00   Median :  140   Median :43.50   Median :  26.00   Median :93.00   Median : 5.800    Median :93.00   Median : 0.100   Median :  2808.74
 Mean   :79.06   Mean   : 2886   Mean   :38.38   Mean   :  56.25   Mean   :82.67   Mean   : 5.938    Mean   :82.48   Mean   : 1.738   Mean   :  7429.17
 3rd Qu.:96.00   3rd Qu.: 2372   3rd Qu.:56.10   3rd Qu.:  56.25   3rd Qu.:97.00   3rd Qu.: 7.452    3rd Qu.:97.00   3rd Qu.: 0.700   3rd Qu.:  7429.17
 Max.   :99.00   Max.   :212183  Max.   :77.60   Max.   :2500.00   Max.   :99.00   Max.   :17.600    Max.   :99.00   Max.   :50.600   Max.   :119172.74

 thinness..1.19.years thinness.5.9.years Income.composition.of.resources   Schooling
 Min.   : 0.10        Min.   : 0.100     Min.   :0.2530                  Min.   : 2.80
 1st Qu.: 1.60        1st Qu.: 1.600     1st Qu.:0.5420                  1st Qu.:10.38
 Median : 3.30        Median : 3.350     Median :0.6770                  Median :12.20
 Mean   : 4.84        Mean   : 4.871     Mean   :0.6575                  Mean   :12.14
 3rd Qu.: 7.20        3rd Qu.: 7.200     3rd Qu.:0.7760                  3rd Qu.:14.20
 Max.   :27.70        Max.   :28.600     Max.   :0.9480                  Max.   :20.70
```

```
> newdata$Status<- revalue(newdata$Status, c("Developing"="0"))
> newdata$Status<- revalue(newdata$Status, c("Developed"="1"))
> newdata$Status
   [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  [77] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0
 [153] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [229] 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [305] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [381] 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [457] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [533] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [609] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 [685] 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [761] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [837] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [913] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
 [989] 1 1 1 1 0 0 0 0 0 0 0 0
 [ reached getOption("max.print") -- omitted 1888 entries ]
Levels: 1 0
```

```
> # declation of variables(newdata)
> cAf = newdata$CountryAfrica
> cAm = newdata$CountryAmericas
> cAs = newdata$CountryAsia
> cEu = newdata$CountryEurope
> cOc= newdata$CountryOceania
> yr = newdata$Year
> sts = newdata$Status
> lfey = newdata$Life.expectancy
> atmy = newdata$Adult.Mortality
> itdth = newdata$infant.deaths
> achl = newdata$Alcohol
> perex = newdata$percentage.expenditure
> hb = newdata$Hepatitis.B
> msls = newdata$Measles
> bmi = newdata$BMI
> ufdth = newdata$under.five.deaths
> plio = newdata$Polio
> totalex = newdata$Total.expenditure
> dpria = newdata$Diphtheria
> hiv = newdata$HIV.AIDS
> gdp = newdata$GDP
> th119 = newdata$thinness..1.19.years
> th59 = newdata$thinness.5.9.years
> iccr = newdata$Income.composition.of.resources
> sch = newdata$Schooling
> # correlation > 0.7 : Strong linear relationship
> # 0.3 < correlation < 0.7 : linear relationship
```

```
> columns = cbind(yr,lfey,atmy,itdth,achl,perex,hb,msls,bmi,ufdth,plio,totalex,dpria,hiv,gdp,th119,th59,iccr,sch)
> cor(columns)
                 yr       lfey        atmy        itdth        achl       perex          hb        msls        bmi       ufdth        plio      totalex
yr       1.00000000  0.1713633 -0.08046149 -0.036660646 -0.04409502  0.07343916  0.13906219 -0.08023290  0.1052408 -0.04227293  0.09656002  0.0806236299
lfey     0.17136333  1.0000000 -0.69000959 -0.143780005  0.40825125  0.40716751  0.28516252 -0.15572583  0.5653220 -0.17139768  0.46327793  0.2178426332
atmy    -0.08046149 -0.6900096  1.00000000  0.037328804 -0.19849889 -0.25257897 -0.15770458  0.03048624 -0.3846725  0.05341650 -0.26955999 -0.1150939998
itdth   -0.03666065 -0.1437800  0.03732880  1.000000000 -0.07665598 -0.05501589 -0.20712259  0.48433546 -0.1889597  0.99588241 -0.14525196 -0.1149345265
achl    -0.04409502  0.4082513 -0.19849889 -0.076655979  1.00000000  0.35679914  0.08750888 -0.04972663  0.3342776 -0.07664250  0.22663904  0.3075537984
perex    0.07343916  0.4071675 -0.25257897 -0.055015893  0.35679914  1.00000000  0.01718558 -0.06281219  0.2515544 -0.05714137  0.16067332  0.1927575463
hb       0.13906219  0.2851625 -0.15770458 -0.207122594  0.08750888  0.01718558  1.00000000 -0.13492542  0.1846717 -0.21721298  0.48775959  0.0919437378
msls    -0.08023290 -0.1557258  0.03048624  0.484335458 -0.04972663 -0.06281219 -0.13492542  1.00000000 -0.1749078  0.49129271 -0.13287492 -0.1021348873
bmi      0.10524080  0.5653220 -0.38467252 -0.188959703  0.33427764  0.25155436  0.18467175 -0.17490785  1.0000000 -0.20066357  0.28584812  0.2361188380
ufdth   -0.04227293 -0.1713977  0.05341650  0.995882413 -0.07664250 -0.05714137 -0.21721298  0.49129271 -0.2006636  1.00000000 -0.16433540 -0.1168371614
plio     0.09656002  0.4632779 -0.26955999 -0.145251963  0.22663904  0.16067332  0.48775959 -0.13287492  0.2858481 -0.16433540  1.00000000  0.1405851152
totalex  0.08062363  0.2178426 -0.11509400 -0.114934527  0.30755380  0.19275755  0.09194374 -0.10213489  0.2361188 -0.11683716  0.14058512  1.0000000000
dpria    0.13807355  0.4769532 -0.26971617 -0.153247383  0.21976750  0.15368871  0.58558458 -0.13754921  0.2833719 -0.17411483  0.67218029  0.1562434948
hiv     -0.13912758 -0.5590283  0.52560264 -0.003479042 -0.05008910 -0.11527353 -0.12748638  0.02247074 -0.2448810  0.01041426 -0.15894470 -0.0005397796
gdp      0.09496966  0.4390118 -0.28241628 -0.064830223  0.32934962  0.90485744  0.06285123 -0.07335667  0.2831931 -0.06846201  0.20253152  0.1232578027
th119   -0.04671259 -0.4775712  0.30292252  0.431672710 -0.43371489 -0.26931280 -0.14127787  0.21504835 -0.5315203  0.43450339 -0.22429171 -0.2801158748
th59    -0.04975056 -0.4720338  0.30864853  0.437009486 -0.42298478 -0.27199079 -0.14360482  0.21103720 -0.5385247  0.43847137 -0.22502480 -0.2878653547
iccr     0.14282444  0.8525520 -0.54627582 -0.142458584  0.52677904  0.47311425  0.25114354 -0.12617168  0.5958531 -0.16396142  0.43832996  0.1795170714
sch      0.19105951  0.7514156 -0.44536169 -0.160569914  0.54713113  0.43219934  0.23920243 -0.13780351  0.5546153 -0.17826766  0.41127781  0.2534972122
             dpria           hiv         gdp       th119        th59       iccr        sch
yr       0.1380735 -0.1391275786  0.09496966 -0.04671259 -0.04975056  0.1428244  0.1910595
lfey     0.4769532 -0.5590283102  0.43901178 -0.47757116 -0.47203384  0.8525520  0.7514156
atmy    -0.2697162  0.5256026449 -0.28241628  0.30292252  0.30864853 -0.5462758 -0.4453617
itdth   -0.1532474 -0.0034790419 -0.06483022  0.43167271  0.43700949 -0.1424586 -0.1605699
achl     0.2197675 -0.0500891036  0.32934962 -0.43371489 -0.42298478  0.5267790  0.5471311
perex    0.1536887 -0.1152735336  0.90485744 -0.26931280 -0.27199079  0.4731142  0.4321993
hb       0.5855846 -0.1274863834  0.06285123 -0.14127787 -0.14360482  0.2511435  0.2392024
msls    -0.1375492  0.0224707445 -0.07335667  0.21504835  0.21103720 -0.1261717 -0.1378035
bmi      0.2833719 -0.2448809564  0.28319309 -0.53152033 -0.53852471  0.5958531  0.5546153
```

```
> columns3 = cbind(y,yr,itdth,hb,msls,ufdth,totalex)
> cor(columns3)
                y          yr       itdth          hb       msls       ufdth     totalex
y       1.0000000  0.17136333 -0.14378001  0.28516252 -0.1557258 -0.17139768  0.21784263
yr      0.1713633  1.00000000 -0.03666065  0.13906219 -0.0802329 -0.04227293  0.08062363
itdth  -0.1437800 -0.03666065  1.00000000 -0.20712259  0.4843355  0.99588241 -0.11493453
hb      0.2851625  0.13906219 -0.20712259  1.00000000 -0.1349254 -0.21721298  0.09194374
msls   -0.1557258 -0.08023290  0.48433546 -0.13492542  1.0000000  0.49129271 -0.10213489
ufdth  -0.1713977 -0.04227293  0.99588241 -0.21721298  0.4912927  1.00000000 -0.11683716
totalex 0.2178426  0.08062363 -0.11493453  0.09194374 -0.1021349 -0.11683716  1.00000000
> #comparison yr tranasformation
> t1 = yr
> t2 = yr*yr
> t3 = log(yr)
> t4 = 1/yr
> t5 = sqrt(yr)
> tcol = cbind(y,t1,t2,t3,t4,t5)
> cor(tcol) # <- No significant improvement, so ignore yr.
           y         t1         t2         t3         t4         t5
y   1.0000000  0.1713633  0.1713643  0.1713621 -0.1713608  0.1713628
t1  0.1713633  1.0000000  0.9999995  0.9999995 -0.9999979  0.9999999
t2  0.1713643  0.9999995  1.0000000  0.9999979 -0.9999953  0.9999988
t3  0.1713621  0.9999995  0.9999979  1.0000000 -0.9999995  0.9999999
t4 -0.1713608 -0.9999979 -0.9999953 -0.9999995  1.0000000 -0.9999988
t5  0.1713628  0.9999999  0.9999988  0.9999999 -0.9999988  1.0000000

> #comparison itdth tranasformation
> t1 = itdth
> t2 = itdth*itdth
> t3 = log(itdth)
> t4 = 1/itdth
> t5 = sqrt(itdth)
> tcol = cbind(y,t1,t2,t3,t4,t5)
> cor(tcol) # <- No significant improvement, so ignore itdth
            y         t1          t2         t3          t4         t5
y   1.00000000 -0.1437800 -0.06143599 -0.2601978  0.26754702 -0.2293294
t1 -0.14378001  1.0000000  0.91196960  0.4966566 -0.22273929  0.8648520
t2 -0.06143599  0.9119696  1.00000000  0.2542002 -0.06817991  0.6332025
t3 -0.26019778  0.4966566  0.25420017  1.0000000 -0.86931860  0.8432264
t4  0.26754702 -0.2227393 -0.06817991 -0.8693186  1.00000000 -0.5570586
t5 -0.22932937  0.8648520  0.63320254  0.8432264 -0.55705863  1.0000000

> #comparison hb tranasformation
> t1 = hb
> t2 = hb*hb
> t3 = log(hb)
> t4 = 1/hb
> t5 = sqrt(hb)
> tcol = cbind(y,t1,t2,t3,t4,t5)
> cor(tcol) # <- significant improvement in t2
            y         t1         t2         t3          t4         t5
y   1.00000000  0.2851625  0.3347489  0.1908003 -0.09164908  0.2428500
t1  0.28516252  1.0000000  0.9733375  0.9426095 -0.74360109  0.9872854
t2  0.33474887  0.9733375  1.0000000  0.8465594 -0.61533884  0.9257378
t3  0.19080034  0.9426095  0.8465594  1.0000000 -0.90050204  0.9831866
t4 -0.09164908 -0.7436011 -0.6153388 -0.9005020  1.00000000 -0.8222637
t5  0.24284996  0.9872854  0.9257378  0.9831866 -0.82226373  1.0000000
```

```
> #comparison msls tranasformation
> t1 = msls
> t2 = msls*msls
> t3 = log(msls)
> t5 = sqrt(msls)
> t4 = 1/msls
> tcol = cbind(y,t1,t2,t3,t4,t5)
> cor(tcol) # <- No significant improvement, so ignore msls.
            y         t1          t2         t3          t4         t5
y   1.00000000 -0.1557258 -0.09991608 -0.2870027  0.16464654 -0.2302218
t1 -0.15572583  1.0000000  0.88211945  0.4312319 -0.11340731  0.8716766
t2 -0.09991608  0.8821195  1.00000000  0.2130945 -0.04327625  0.6087597
t3 -0.28700268  0.4312319  0.21309448  1.0000000 -0.66650168  0.7562982
t4  0.16464654 -0.1134073 -0.04327625 -0.6665017  1.00000000 -0.2847730
t5 -0.23022178  0.8716766  0.60875972  0.7562982 -0.28477303  1.0000000
> #comparison ufdth tranasformation
> t1 = ufdth
> t2 = ufdth*ufdth
> t3 = log(ufdth)
> t4 = 1/ufdth
> t5 = sqrt(ufdth)
> tcol = cbind(y,t1,t2,t3,t4,t5)
> cor(tcol) # <- significant improvement in t3
            y         t1          t2         t3          t4         t5
y   1.00000000 -0.1713977 -0.07161273 -0.3011642  0.28855512 -0.2725487
t1 -0.17139768  1.0000000  0.90981559  0.4920628 -0.21729998  0.8627050
t2 -0.07161273  0.9098156  1.00000000  0.2482720 -0.06662786  0.6279305
t3 -0.30116422  0.4920628  0.24827205  1.0000000 -0.85637568  0.8413576
t4  0.28855512 -0.2173000 -0.06662786 -0.8563757  1.00000000 -0.5422894
t5 -0.27254873  0.8627050  0.62793052  0.8413576 -0.54228943  1.0000000

> #comparison totalex tranasformation
> t1 = totalex
> t2 = totalex*totalex
> t3 = log(totalex)
> t4 = 1/totalex
> t5 = sqrt(totalex)
> tcol = cbind(y,t1,t2,t3,t4,t5)
> cor(tcol) # <- No significant improvement, so ignore totalex
            y         t1          t2         t3          t4         t5
y   1.00000000  0.2178426  0.2257854  0.1576599 -0.06547376  0.1942793
t1  0.21784263  1.0000000  0.9564450  0.9422587 -0.74419499  0.9865021
t2  0.22578545  0.9564450  1.0000000  0.8155386 -0.57618002  0.8989354
t3  0.15765986  0.9422587  0.8155386  1.0000000 -0.91024429  0.9838098
t4 -0.06547376 -0.7441950 -0.5761800 -0.9102443  1.00000000 -0.8303562
t5  0.19427927  0.9865021  0.8989354  0.9838098 -0.83035616  1.0000000
> ###Conclusion: (ignore yr,msls,totalex), (transform sqrt(itdth), hb2, log(ufdth))
> newdata = newdata[c(-6,-10,-14,-18)] #(ignore yr,itdth,msls,totalex)
> #add transformed new variables
> hb2 = hb*hb
> newdata[,"hb2"]=hb2
> logufdth = log(ufdth)
> newdata[,"logufdth"]=logufdth
```

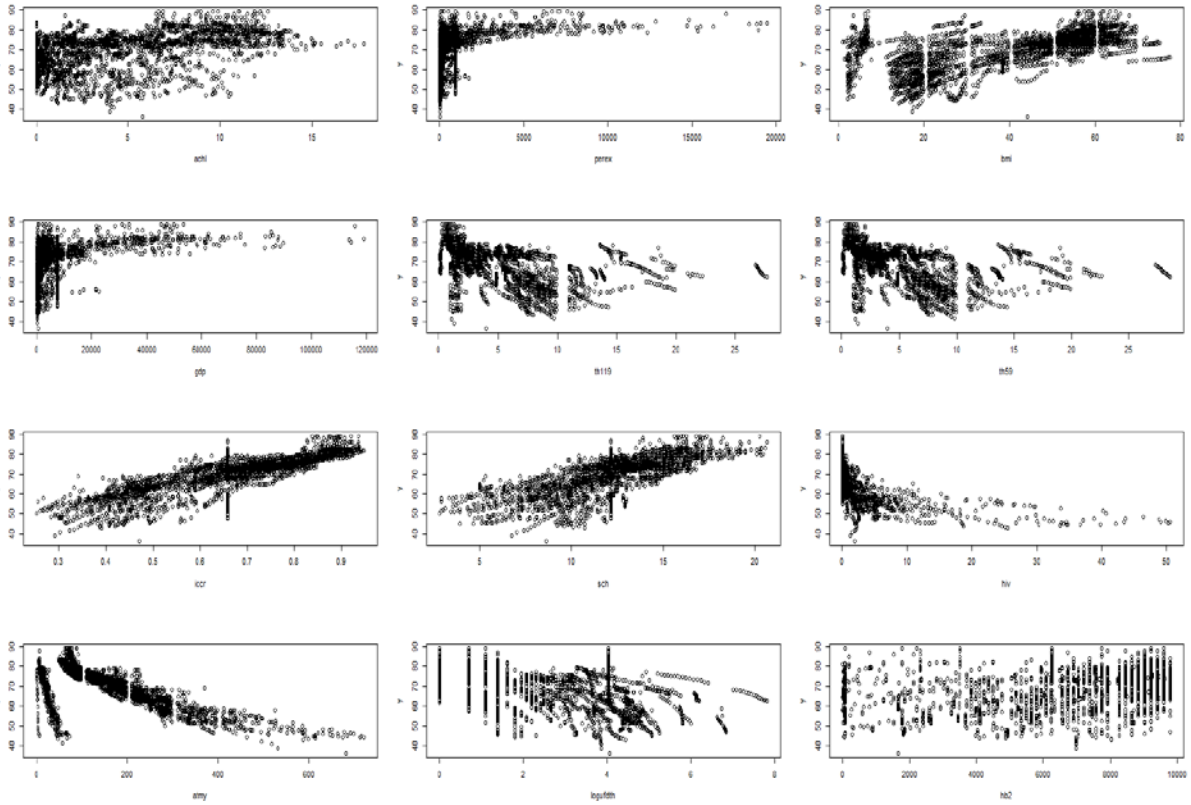
```
> library(car)
> #build basic model(Just only one x variables)
> basicfull = glm(lfey~cAf + cAm+ cAs+cOc+
+             sts+
+             achl+perex+bmi+gdp+th119+th59+iccr+sch+hiv+
+             atmy+logufdth+hb2+plio+dpria)
> # Evaluate Multicollinearity
> vif(basicfull) #iccr>th59>th119>sch>gdp>perex>cAf>
      cAf       cAm       cAs       cOc       sts      achl     perex       bmi       gdp     th119      th59      iccr       sch       hiv      atmy
 5.036362  2.384928  3.921990  1.679143  2.417647  2.586669  6.058090  1.953509  6.202818  8.747050  8.918560 11.108005  7.700737  1.583615  1.826753
 logufdth       hb2      plio     dpria
 1.235896  1.782178  2.016860  2.313377
> basicfull = glm(lfey~ cAm+ cAs+cOc+ #iccr,th59,perex,cAf,cEu are removed.
+                 sts+
+                 achl+bmi+gdp+th119+sch+hiv+
+                 atmy+logufdth+hb2+plio+dpria)
> vif(basicfull)
     cAm      cAs      cOc      sts     achl      bmi      gdp     th119      sch      hiv     atmy logufdth      hb2     plio    dpria
1.487572 1.633895 1.271044 2.225988 2.261671 1.771586 1.414127 1.916015 2.492591 1.461849 1.762092 1.224861 1.739596 2.005159 2.299622
> #build full model CASE1
> full_1 = glm(lfey~ cAm+ cAs+cOc+ #iccr,th59,perex,cAf,cEu are removed.
+                 sts+
+                 achl+bmi+gdp+th119+sch+hiv+
+                 atmy+logufdth+hb2+plio+dpria)
> #build base model CASE1
> base_1 = glm(lfey~achl,data=newdata)
```

```
> summary(m1_1)  # aic 15990

Call:
glm(formula = lfey ~ sch + atmy + hiv + dpria + bmi + cAs + cAm +
    sts + gdp + logufdth + plio + th119 + cOc, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.2957   -2.2662    0.0203    2.3531   13.2853

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.610e+01  6.219e-01  90.207  < 2e-16 ***
sch          1.063e+00  3.541e-02  30.011  < 2e-16 ***
atmy        -1.702e-02  7.693e-04 -22.118  < 2e-16 ***
hiv         -4.286e-01  1.690e-02 -25.364  < 2e-16 ***
dpria        3.204e-02  4.236e-03   7.564 5.24e-14 ***
bmi          3.810e-02  4.786e-03   7.962 2.42e-15 ***
cAs          3.062e+00  2.018e-01  15.172  < 2e-16 ***
cAm          3.311e+00  2.240e-01  14.782  < 2e-16 ***
sts0        -3.580e+00  2.544e-01 -14.074  < 2e-16 ***
gdp          4.831e-05  6.400e-06   7.549 5.85e-14 ***
logufdth    -2.885e-01  4.595e-02  -6.278 3.95e-10 ***
plio         2.644e-02  4.255e-03   6.213 5.94e-10 ***
th119       -9.042e-02  2.233e-02  -4.049 5.29e-05 ***
cOc          7.375e-01  3.448e-01   2.139   0.0325 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.80138)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42539  on 2874  degrees of freedom
AIC: 15994

Number of Fisher Scoring iterations: 2
```

```
> summary(m1_2) # AIC: 16000

Call:
glm(formula = lfey ~ achl + sch + atmy + hiv + dpria + bmi +
    cAs + cAm + sts + gdp + logufdth + plio + th119 + cOc, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.3014   -2.2602    0.0272    2.3640   13.2496

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.608e+01  6.304e-01  88.957  < 2e-16 ***
achl         5.895e-03  2.657e-02   0.222   0.8244
sch          1.060e+00  3.700e-02  28.659  < 2e-16 ***
atmy        -1.703e-02  7.705e-04 -22.096  < 2e-16 ***
hiv         -4.288e-01  1.692e-02 -25.343  < 2e-16 ***
dpria        3.206e-02  4.237e-03   7.565 5.18e-14 ***
bmi          3.811e-02  4.787e-03   7.961 2.44e-15 ***
cAs          3.075e+00  2.096e-01  14.667  < 2e-16 ***
cAm          3.305e+00  2.259e-01  14.627  < 2e-16 ***
sts0        -3.554e+00  2.797e-01 -12.707  < 2e-16 ***
gdp          4.830e-05  6.401e-06   7.547 5.95e-14 ***
logufdth    -2.883e-01  4.596e-02  -6.272 4.09e-10 ***
plio         2.642e-02  4.258e-03   6.205 6.27e-10 ***
th119       -8.999e-02  2.242e-02  -4.013 6.14e-05 ***
cOc          7.541e-01  3.529e-01   2.137   0.0327 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.80628)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42538  on 2873  degrees of freedom
AIC: 15996

Number of Fisher Scoring iterations: 2
```

```
> summary(m1_3) # AIC: 15990

Call:
glm(formula = lfey ~ cAm + cAs + cOc + sts + bmi + gdp + th119 +
    sch + hiv + atmy + logufdth + plio + dpria)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.2957   -2.2662    0.0203    2.3531   13.2853

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.610e+01  6.219e-01  90.207  < 2e-16 ***
cAm          3.311e+00  2.240e-01  14.782  < 2e-16 ***
cAs          3.062e+00  2.018e-01  15.172  < 2e-16 ***
cOc          7.375e-01  3.448e-01   2.139   0.0325 *
sts0        -3.580e+00  2.544e-01 -14.074  < 2e-16 ***
bmi          3.810e-02  4.786e-03   7.962 2.42e-15 ***
gdp          4.831e-05  6.400e-06   7.549 5.85e-14 ***
th119       -9.042e-02  2.233e-02  -4.049 5.29e-05 ***
sch          1.063e+00  3.541e-02  30.011  < 2e-16 ***
hiv         -4.286e-01  1.690e-02 -25.364  < 2e-16 ***
atmy        -1.702e-02  7.693e-04 -22.118  < 2e-16 ***
logufdth    -2.885e-01  4.595e-02  -6.278 3.95e-10 ***
plio         2.644e-02  4.255e-03   6.213 5.94e-10 ***
dpria        3.204e-02  4.236e-03   7.564 5.24e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.80138)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42539  on 2874  degrees of freedom
AIC: 15994

Number of Fisher Scoring iterations: 2
```

```
> summary(m1_4) # AIC - 16000

Call:
glm(formula = lfey ~ cAm + cAs + cOc + sts + achl + bmi + gdp +
    th119 + sch + hiv + atmy + logufdth + hb2 + plio + dpria,
    data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.4106   -2.2666   -0.0036    2.3771   13.1216

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.607e+01  6.305e-01  88.942  < 2e-16 ***
cAm          3.307e+00  2.259e-01  14.636  < 2e-16 ***
cAs          3.085e+00  2.100e-01  14.689  < 2e-16 ***
cOc          7.510e-01  3.529e-01   2.128   0.0334 *
sts0        -3.551e+00  2.797e-01 -12.696  < 2e-16 ***
achl         5.289e-03  2.659e-02   0.199   0.8423
bmi          3.820e-02  4.788e-03   7.978 2.13e-15 ***
gdp          4.788e-05  6.422e-06   7.455 1.18e-13 ***
th119       -9.031e-02  2.243e-02  -4.027 5.80e-05 ***
sch          1.061e+00  3.702e-02  28.670  < 2e-16 ***
hiv         -4.293e-01  1.693e-02 -25.355  < 2e-16 ***
atmy        -1.701e-02  7.708e-04 -22.064  < 2e-16 ***
logufdth    -2.911e-01  4.609e-02  -6.316 3.10e-10 ***
hb2         -2.718e-05  3.272e-05  -0.831   0.4063
plio         2.717e-02  4.352e-03   6.242 4.97e-10 ***
dpria        3.354e-02  4.598e-03   7.294 3.87e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.80788)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42528  on 2872  degrees of freedom
AIC: 15997

Number of Fisher Scoring iterations: 2
```

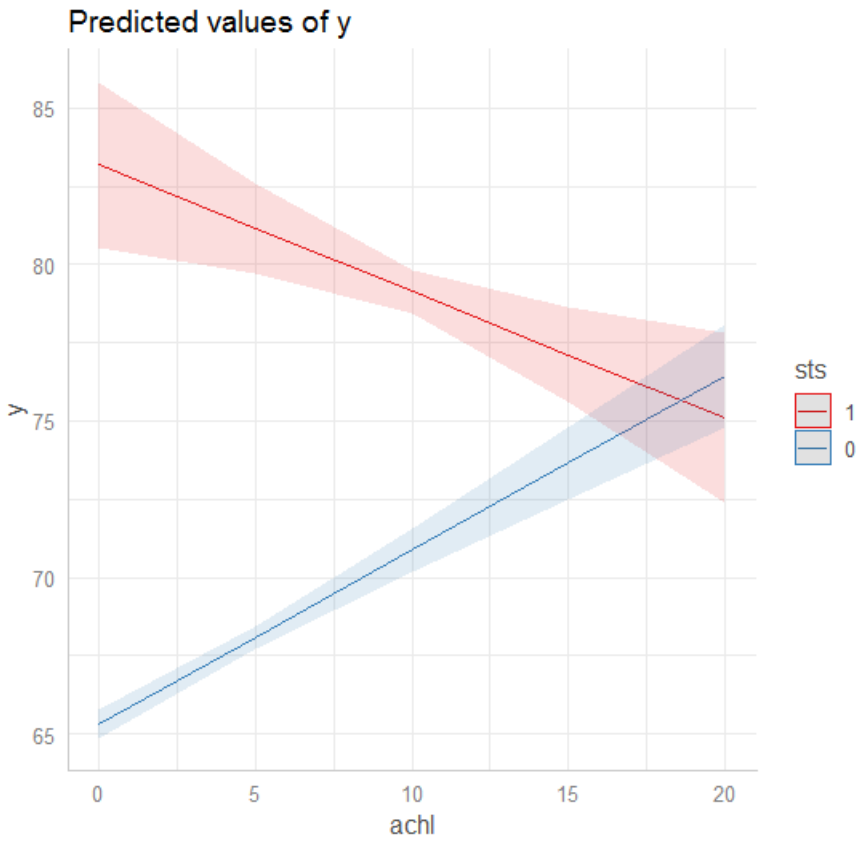


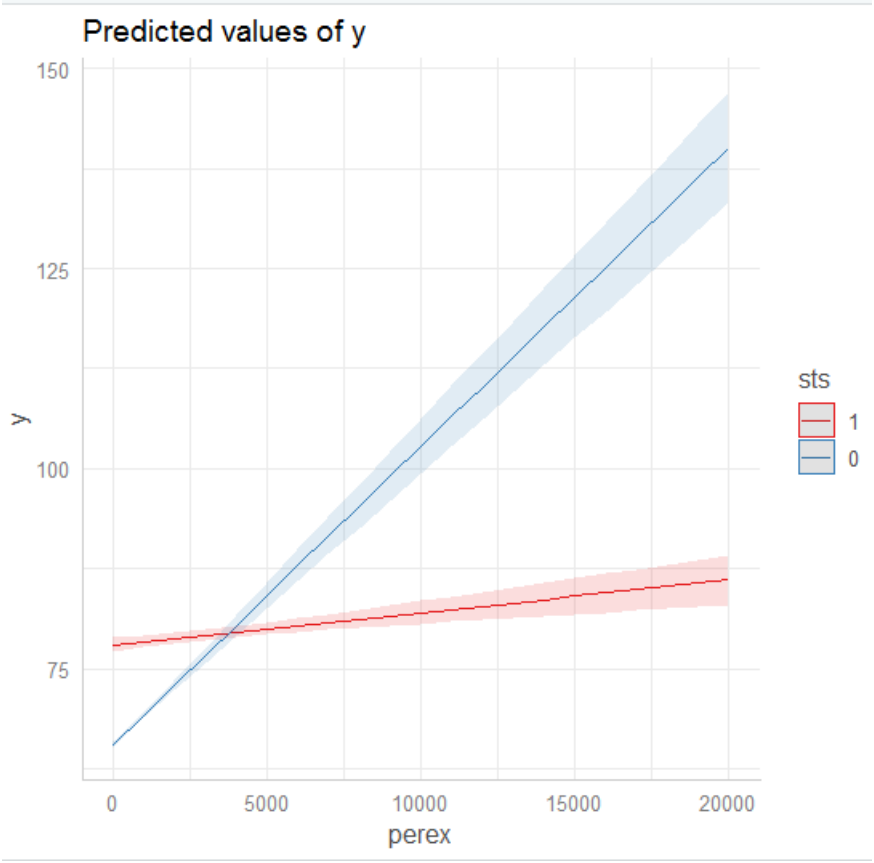
```
> full_2 = glm(lfey~cAm+cAs+cOc+
+               sts+
+               achl+bmi+gdp+th119+th1192+th1193+
+               sch+hiv+atmy+logufdth+logufdth2+logufdth3+hb2+plio+dpria,
+             data=newdata)
> vif(full_2)
      cAm       cAs       cOc       sts      achl       bmi       gdp      th119     th1192     th1193       sch       hiv      atmy   logufdth
  1.592216  1.716524  1.483738  2.429498  2.363076  1.801535  1.425913 42.673789 175.131407  75.088874  2.498266  1.482562  1.776987 49.754289
 logufdth2 logufdth3       hb2      plio     dpria
237.755137 99.989116  1.828378  2.014751  2.327279
> full_2 = glm(lfey~cAm+cAs+cOc+
+               sts+
+               achl+bmi+gdp+th1193+
+               sch+hiv+atmy+logufdth+hb2+plio+dpria,
+             data=newdata)
> vif(full_2) #Multicollinearity 해결
     cAm      cAs      cOc      sts     achl      bmi      gdp   th1193      sch      hiv     atmy logufdth      hb2     plio    dpria
1.464359 1.677329 1.223138 2.207279 2.244920 1.694000 1.413451 1.293013 2.477531 1.452164 1.759284 1.242929 1.766383 2.005792 2.299238
> #build base model CASE1
> base_2 = glm(lfey~achl,data=newdata)
> # model 1  Both direction by stepwise()
> m2_1 = step(base_2, scope=list(upper=full_2, lower=~1), direction="both", trace=T)
```

```
> summary(m2_1)  # aic 16010

Call:
glm(formula = lfey ~ sch + atmy + hiv + dpria + bmi + cAs + cAm +
    sts + gdp + logufdth + plio + cOc, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.1784   -2.3435    0.0455    2.4088   13.6174

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.547e+01  6.039e-01  91.859  < 2e-16 ***
sch          1.080e+00  3.524e-02  30.653  < 2e-16 ***
atmy        -1.713e-02  7.709e-04 -22.223  < 2e-16 ***
hiv         -4.341e-01  1.689e-02 -25.701  < 2e-16 ***
dpria        3.267e-02  4.245e-03   7.696 1.91e-14 ***
bmi          4.283e-02  4.654e-03   9.202  < 2e-16 ***
cAs          2.841e+00  1.948e-01  14.583  < 2e-16 ***
cAm          3.439e+00  2.224e-01  15.463  < 2e-16 ***
sts0        -3.729e+00  2.524e-01 -14.775  < 2e-16 ***
gdp          4.988e-05  6.405e-06   7.788 9.45e-15 ***
logufdth    -3.237e-01  4.523e-02  -7.157 1.04e-12 ***
plio         2.644e-02  4.267e-03   6.198 6.56e-10 ***
cOc          9.899e-01  3.400e-01   2.912  0.00362 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.88063)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42782  on 2875  degrees of freedom
AIC: 16009

Number of Fisher Scoring iterations: 2
```

```
> summary(m2_2) # AIC: 16010

Call:
glm(formula = lfey ~ achl + sch + atmy + hiv + dpria + bmi +
    cAs + cAm + sts + gdp + logufdth + plio + cOc, data = newdata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-20.195   -2.337    0.050    2.415   13.507

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.542e+01  6.105e-01  90.788  < 2e-16 ***
achl         1.520e-02  2.654e-02   0.573  0.56696
sch          1.074e+00  3.694e-02  29.071  < 2e-16 ***
atmy        -1.715e-02  7.719e-04 -22.222  < 2e-16 ***
hiv         -4.345e-01  1.691e-02 -25.701  < 2e-16 ***
dpria        3.270e-02  4.246e-03   7.702 1.83e-14 ***
bmi          4.277e-02  4.656e-03   9.188  < 2e-16 ***
cAs          2.876e+00  2.042e-01  14.081  < 2e-16 ***
cAm          3.420e+00  2.247e-01  15.224  < 2e-16 ***
sts0        -3.660e+00  2.792e-01 -13.112  < 2e-16 ***
gdp          4.984e-05  6.406e-06   7.781 9.98e-15 ***
logufdth    -3.229e-01  4.526e-02  -7.133 1.24e-12 ***
plio         2.638e-02  4.269e-03   6.180 7.31e-10 ***
cOc          1.030e+00  3.470e-01   2.967  0.00303 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.8841)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42777  on 2874  degrees of freedom
AIC: 16010

Number of Fisher Scoring iterations: 2


> summary(m2_3) # AIC: 16010

Call:
glm(formula = lfey ~ cAm + cAs + cOc + sts + bmi + gdp + sch +
    hiv + atmy + logufdth + plio + dpria, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.1784   -2.3435    0.0455    2.4088   13.6174

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.547e+01  6.039e-01  91.859  < 2e-16 ***
cAm          3.439e+00  2.224e-01  15.463  < 2e-16 ***
cAs          2.841e+00  1.948e-01  14.583  < 2e-16 ***
cOc          9.899e-01  3.400e-01   2.912  0.00362 **
sts0        -3.729e+00  2.524e-01 -14.775  < 2e-16 ***
bmi          4.283e-02  4.654e-03   9.202  < 2e-16 ***
gdp          4.988e-05  6.405e-06   7.788 9.45e-15 ***
sch          1.080e+00  3.524e-02  30.653  < 2e-16 ***
hiv         -4.341e-01  1.689e-02 -25.701  < 2e-16 ***
atmy        -1.713e-02  7.709e-04 -22.223  < 2e-16 ***
logufdth    -3.237e-01  4.523e-02  -7.157 1.04e-12 ***
plio         2.644e-02  4.267e-03   6.198 6.56e-10 ***
dpria        3.267e-02  4.245e-03   7.696 1.91e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.88063)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42782  on 2875  degrees of freedom
AIC: 16009

Number of Fisher Scoring iterations: 2
```

```
> summary(m2_4) # AIC - 16010

Call:
glm(formula = lfey ~ cAm + cAs + cOc + sts + bmi + gdp + sch +
    hiv + atmy + logufdth + plio + dpria, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.1784   -2.3435    0.0455    2.4088   13.6174

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.547e+01  6.039e-01  91.859  < 2e-16 ***
cAm          3.439e+00  2.224e-01  15.463  < 2e-16 ***
cAs          2.841e+00  1.948e-01  14.583  < 2e-16 ***
cOc          9.899e-01  3.400e-01   2.912  0.00362 **
sts0        -3.729e+00  2.524e-01 -14.775  < 2e-16 ***
bmi          4.283e-02  4.654e-03   9.202  < 2e-16 ***
gdp          4.988e-05  6.405e-06   7.788 9.45e-15 ***
sch          1.080e+00  3.524e-02  30.653  < 2e-16 ***
hiv         -4.341e-01  1.689e-02 -25.701  < 2e-16 ***
atmy        -1.713e-02  7.709e-04 -22.223  < 2e-16 ***
logufdth    -3.237e-01  4.523e-02  -7.157 1.04e-12 ***
plio         2.644e-02  4.267e-03   6.198 6.56e-10 ***
dpria        3.267e-02  4.245e-03   7.696 1.91e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 14.88063)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  42782  on 2875  degrees of freedom
AIC: 16009

Number of Fisher Scoring iterations: 2
```





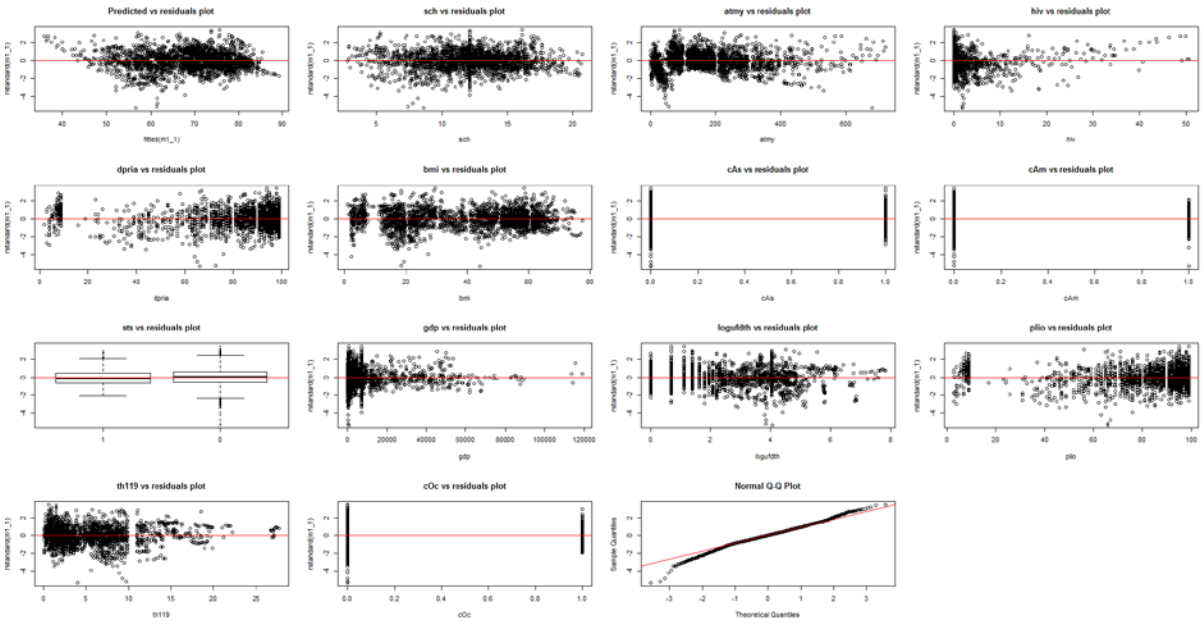

```
> vif(full_3)
    achl      bmi      gdp     th119      sch      hiv     atmy logufdth      hb2     plio     dpria      cAm      cAs      cOc  gdp:cAm
2.206891 1.841391 2.041628 2.095854 2.557634 1.474745 1.737227 1.250377 1.821555 2.009867 2.315663 2.588659 3.554218 2.932791 1.849678
 hiv:cAm achl:cAs  gdp:cAs  hiv:cAs  gdp:cOc th119:cOc  hiv:cOc
1.399130 1.915873 1.847927 1.537586 1.449417 1.780363 1.453471
> max(vif(full_3)) # all the multicollinearity variables are removed.
[1] 3.554218
```

```
> summary(m3_1)  # aic 16124

Call:
glm(formula = lfey ~ achl + sch + atmy + hiv + dpria + bmi + 
    cAs + cAm + gdp + logufdth + plio + th119 + cOc + hiv:cAm + 
    cAm:gdp + cAs:gdp + gdp:cOc + th119:cOc, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max  
-20.3336   -2.1903    0.0896    2.3774   12.9885  

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  5.213e+01  5.681e-01  91.775  < 2e-16 ***
achl         1.457e-01  2.506e-02   5.813 6.83e-09 ***
sch          1.120e+00  3.796e-02  29.519  < 2e-16 ***
atmy        -1.851e-02  7.801e-04 -23.722  < 2e-16 ***
hiv         -4.352e-01  1.734e-02 -25.093  < 2e-16 ***
dpria        3.297e-02  4.334e-03   7.608 3.75e-14 ***
bmi          3.697e-02  4.914e-03   7.523 7.14e-14 ***
cAs          3.001e+00  2.344e-01  12.803  < 2e-16 ***
cAm          2.972e+00  3.023e-01   9.830  < 2e-16 ***
gdp          8.324e-05  7.765e-06  10.721  < 2e-16 ***
logufdth    -2.833e-01  4.726e-02  -5.994 2.30e-09 ***
plio         2.586e-02  4.357e-03   5.934 3.31e-09 ***
th119       -1.060e-01  2.310e-02  -4.589 4.64e-06 ***
cOc          2.041e+00  5.078e-01   4.020 5.96e-05 ***
hiv:cAm     -7.642e-01  2.628e-01  -2.908  0.00366 ** 
cAm:gdp     -7.589e-05  2.751e-05  -2.758  0.00585 ** 
cAs:gdp     -3.483e-05  1.364e-05  -2.554  0.01070 *  
gdp:cOc     -1.079e-04  2.438e-05  -4.426 9.98e-06 ***
th119:cOc   -6.447e-01  2.721e-01  -2.370  0.01787 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 15.45665)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  44345  on 2869  degrees of freedom
AIC: 16124

Number of Fisher Scoring iterations: 2
```

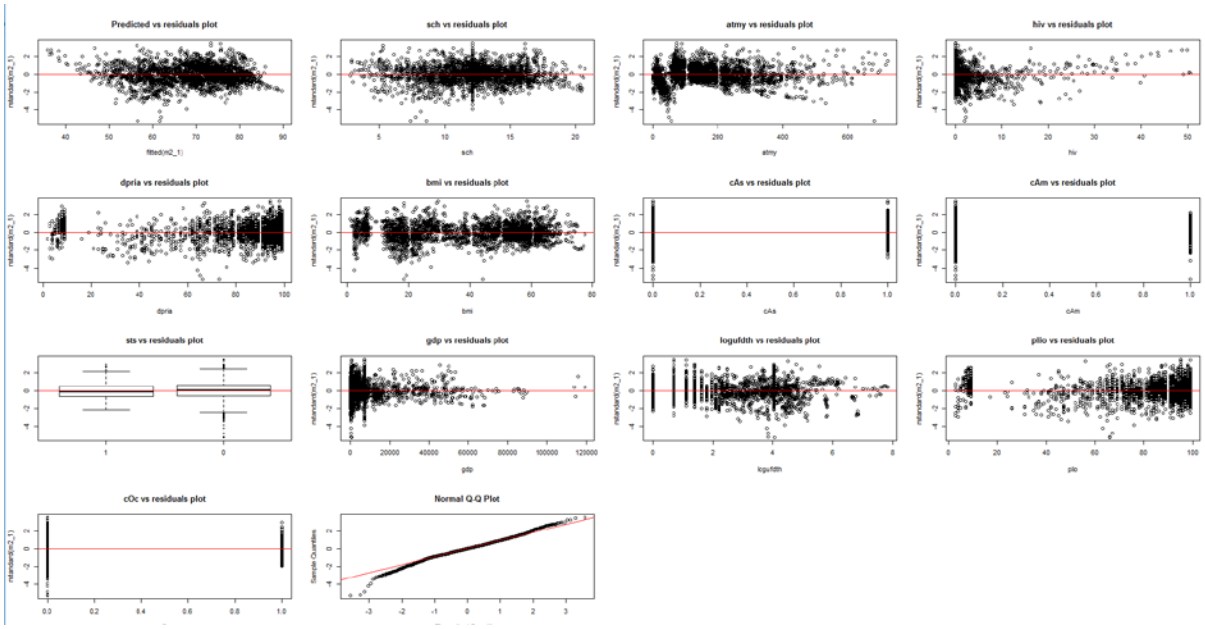
```
> summary(m3_2) # AIC: 16124

Call:
glm(formula = lfey ~ achl + sch + atmy + hiv + dpria + bmi +
    cAs + cAm + gdp + logufdth + plio + th119 + cOc + hiv:cAm +
    cAm:gdp + cAs:gdp + gdp:cOc + th119:cOc, data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.3336   -2.1903    0.0896    2.3774   12.9885

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.213e+01  5.681e-01  91.775  < 2e-16 ***
achl         1.457e-01  2.506e-02   5.813 6.83e-09 ***
sch          1.120e+00  3.796e-02  29.519  < 2e-16 ***
atmy        -1.851e-02  7.801e-04 -23.722  < 2e-16 ***
hiv         -4.352e-01  1.734e-02 -25.093  < 2e-16 ***
dpria        3.297e-02  4.334e-03   7.608 3.75e-14 ***
bmi          3.697e-02  4.914e-03   7.523 7.14e-14 ***
cAs          3.001e+00  2.344e-01  12.803  < 2e-16 ***
cAm          2.972e+00  3.023e-01   9.830  < 2e-16 ***
gdp          8.324e-05  7.765e-06  10.721  < 2e-16 ***
logufdth    -2.833e-01  4.726e-02  -5.994 2.30e-09 ***
plio         2.586e-02  4.357e-03   5.934 3.31e-09 ***
th119       -1.060e-01  2.310e-02  -4.589 4.64e-06 ***
cOc          2.041e+00  5.078e-01   4.020 5.96e-05 ***
hiv:cAm     -7.642e-01  2.628e-01  -2.908  0.00366 **
cAm:gdp     -7.589e-05  2.751e-05  -2.758  0.00585 **
cAs:gdp     -3.483e-05  1.364e-05  -2.554  0.01070 *
gdp:cOc     -1.079e-04  2.438e-05  -4.426 9.98e-06 ***
th119:cOc   -6.447e-01  2.721e-01  -2.370  0.01787 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 15.45665)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  44345  on 2869  degrees of freedom
AIC: 16124
```

```
> summary(m3_3) # AIC: 16124

Call:
glm(formula = lfey ~ achl + bmi + gdp + th119 + sch + hiv + atmy +
    logufdth + plio + dpria + cAm + cAs + cOc + gdp:cAm + hiv:cAm +
    gdp:cAs + gdp:cOc + th119:cOc)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.3336   -2.1903    0.0896    2.3774   12.9885

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.213e+01  5.681e-01  91.775  < 2e-16 ***
achl         1.457e-01  2.506e-02   5.813 6.83e-09 ***
bmi          3.697e-02  4.914e-03   7.523 7.14e-14 ***
gdp          8.324e-05  7.765e-06  10.721  < 2e-16 ***
th119       -1.060e-01  2.310e-02  -4.589 4.64e-06 ***
sch          1.120e+00  3.796e-02  29.519  < 2e-16 ***
hiv         -4.352e-01  1.734e-02 -25.093  < 2e-16 ***
atmy        -1.851e-02  7.801e-04 -23.722  < 2e-16 ***
logufdth    -2.833e-01  4.726e-02  -5.994 2.30e-09 ***
plio         2.586e-02  4.357e-03   5.934 3.31e-09 ***
dpria        3.297e-02  4.334e-03   7.608 3.75e-14 ***
cAm          2.972e+00  3.023e-01   9.830  < 2e-16 ***
cAs          3.001e+00  2.344e-01  12.803  < 2e-16 ***
cOc          2.041e+00  5.078e-01   4.020 5.96e-05 ***
gdp:cAm     -7.589e-05  2.751e-05  -2.758  0.00585 **
hiv:cAm     -7.642e-01  2.628e-01  -2.908  0.00366 **
gdp:cAs     -3.483e-05  1.364e-05  -2.554  0.01070 *
gdp:cOc     -1.079e-04  2.438e-05  -4.426 9.98e-06 ***
th119:cOc   -6.447e-01  2.721e-01  -2.370  0.01787 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 15.45665)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  44345  on 2869  degrees of freedom
AIC: 16124
```

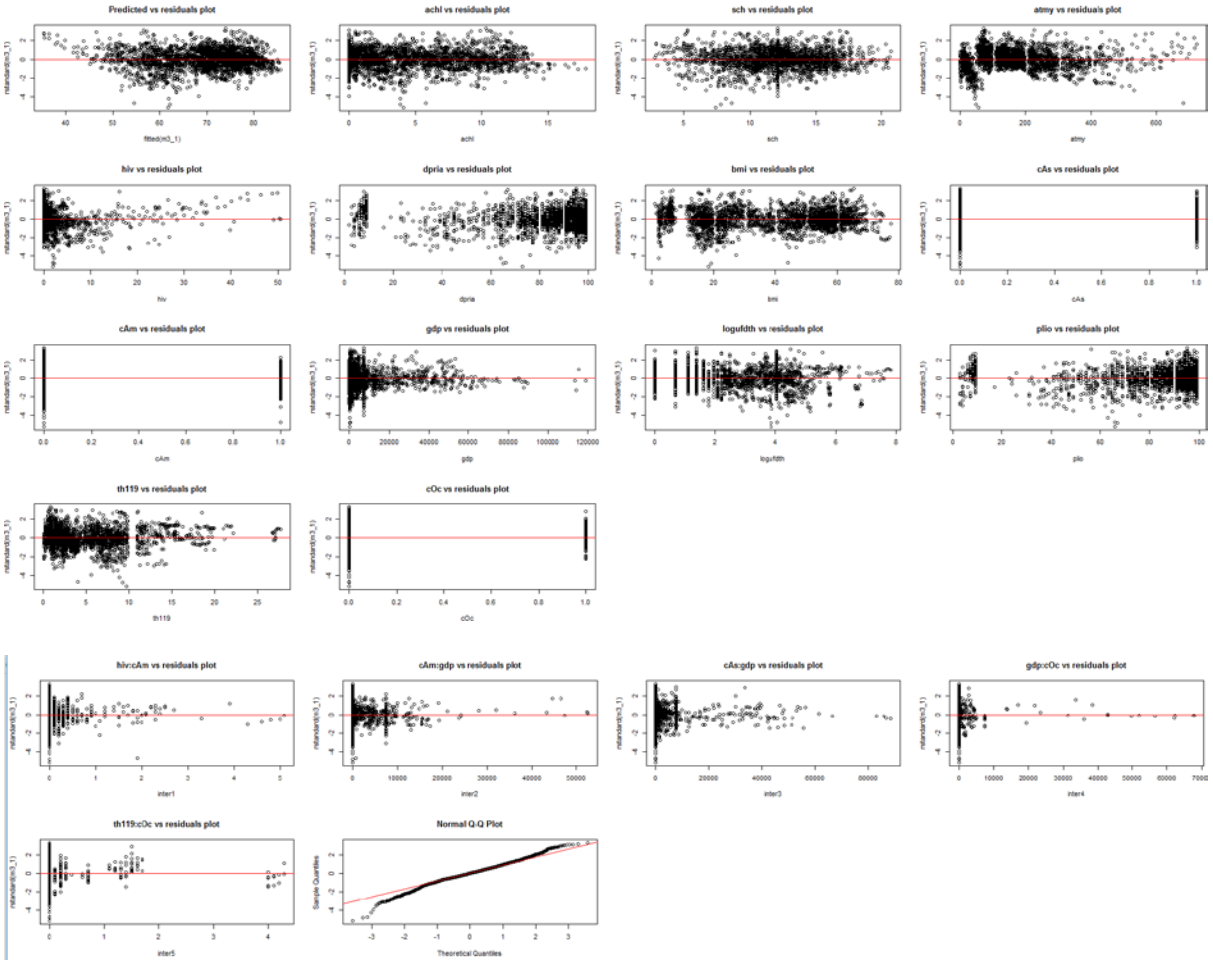
```
> summary(m3_4) # AIC - 16124

Call:
glm(formula = lfey ~ achl + bmi + gdp + th119 + sch + hiv + atmy +
    logufdth + plio + dpria + (gdp * cAm) + (hiv * cAm) + (achl *
    cAs) + (gdp * cAs) + (gdp * cOc) + (th119 * cOc), data = newdata)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-20.3491   -2.1854    0.0923    2.3443   13.0468

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.209e+01  5.689e-01  91.560  < 2e-16 ***
achl         1.328e-01  2.677e-02   4.960 7.45e-07 ***
bmi          3.806e-02  4.977e-03   7.646 2.81e-14 ***
gdp          8.453e-05  7.821e-06  10.809  < 2e-16 ***
th119       -9.790e-02  2.384e-02  -4.107 4.12e-05 ***
sch          1.125e+00  3.809e-02  29.532  < 2e-16 ***
hiv         -4.358e-01  1.735e-02 -25.125  < 2e-16 ***
atmy        -1.853e-02  7.802e-04 -23.751  < 2e-16 ***
logufdth    -2.840e-01  4.726e-02  -6.010 2.09e-09 ***
plio         2.570e-02  4.358e-03   5.897 4.14e-09 ***
dpria        3.292e-02  4.333e-03   7.597 4.07e-14 ***
cAm          2.982e+00  3.024e-01   9.864  < 2e-16 ***
cAs          2.771e+00  2.884e-01   9.609  < 2e-16 ***
cOc          2.018e+00  5.080e-01   3.972 7.31e-05 ***
gdp:cAm     -7.631e-05  2.751e-05  -2.774  0.00558 **
hiv:cAm     -7.644e-01  2.627e-01  -2.909  0.00365 **
achl:cAs     8.699e-02  6.354e-02   1.369  0.17109
gdp:cAs     -3.718e-05  1.374e-05  -2.705  0.00686 **
gdp:cOc     -1.079e-04  2.437e-05  -4.428 9.88e-06 ***
th119:cOc   -6.514e-01  2.721e-01  -2.394  0.01671 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 15.45194)

    Null deviance: 258185  on 2887  degrees of freedom
Residual deviance:  44316  on 2868  degrees of freedom
AIC: 16124

Number of Fisher Scoring iterations: 2
```



Residual Analysis plots of m1_1



Residual Analysis plots of m2_1



Residual Analysis plots of m3_1

```
> errs = cbind(mse1, mse2, mse3)
> errs
          mse1     mse2     mse3
[1,] 167.0548 167.5886 168.6911
[2,] 167.0548 167.5886 168.6911
```

```
> # Removing Outliers
> # influential row numbers
> #install.packages('tidyverse')
> #install.packages('gridExtra')
> library(tidyverse)
> library(gridExtra)
> cooksd <- cooks.distance(m1_1)
> influential <- as.numeric(names(cooksd)[(cooksd > (4/2888))])
> influential
  [1]   49   63   64  113  114  115  116  117  118  120  122  127  133  196  310  317  326  327  334  351  352  410  432  433  435  437  438  439  440  445
 [31]  455  456  483  497  514  517  518  519  520  531  532  533  534  535  536  540  624  625  629  633  666  715  720  721  744  788  840  842  848  884
 [61]  888  889  900  901  903  919  920  922  923  924  961  962  982  983 1012 1025 1043 1087 1089 1090 1100 1112 1156 1222 1223 1233 1364 1365 1366 1367
 [91] 1368 1369 1370 1386 1455 1462 1463 1470 1479 1482 1510 1519 1541 1550 1557 1560 1608 1639 1698 1701 1718 1747 1751 1856 1880 1881 1882 1883 1884 1885
[121] 1893 1937 1961 2033 2034 2066 2067 2068 2069 2070 2075 2084 2140 2176 2183 2185 2186 2189 2190 2275 2276 2277 2278 2279 2280 2281 2282 2283 2284 2285
[151] 2286 2288 2289 2299 2322 2343 2361 2362 2363 2364 2365 2366 2380 2381 2382 2383 2384 2410 2473 2475 2479 2480 2481 2490 2516 2517 2630 2703 2708 2741
[181] 2742 2743 2760 2768 2769 2822 2825 2826 2827 2832 2833 2890 2895 2896 2907 2908 2910 2911 2912 2913 2914
> # Option) Alternatively, we can try to remove the top x outliers to have a look
> # influential <- as.numeric(names(sort(cooksd, decreasing = TRUE)[1:top_x_outlier]))
> newdata2 <- newdata[-influential, ]
> cAm = newdata2$CountryAmericas
> cAs = newdata2$CountryAsia
> cOc= newdata2$CountryOceania
> sts = newdata2$Status
> lfey = newdata2$Life.expectancy
> atmy = newdata2$Adult.Mortality
> bmi = newdata2$BMI
> ufdth = newdata2$under.five.deaths
> plio = newdata2$Polio
> dpria = newdata2$Diphtheria
> hiv = newdata2$HIV.AIDS
> gdp = newdata2$GDP
> th119 = newdata2$thinness..1.19.years
> sch = newdata2$Schooling
```

```
> m1_1_new = glm(lfey ~ sch + atmy + hiv + dpria + bmi + cAs + cAm +
+              sts + gdp + log(newdata2$under.five.deaths) + plio + th119 + cOc, data = newdata2) #에러나서 재설정
> mse1_new = cv.glm(newdata2, m1_1_new, K=10)$delta
There were 20 warnings (use warnings() to see them)
> errs2 = cbind(mse1, mse1_new)
> errs2
        mse1 mse1_new
[1,] 167.0548 157.2787
[2,] 167.0548 157.2787
```

```
> summary(m1_1_new)

Call:
glm(formula = lfey ~ sch + atmy + hiv + dpria + bmi + cAs + cAm +
    sts + gdp + log(newdata2$under.five.deaths) + plio + th119 +
    cOc, data = newdata2)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-19.8298   -2.2155   -0.0393    2.2311   13.4442

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     5.746e+01  6.235e-01  92.147  < 2e-16 ***
sch                             1.003e+00  3.528e-02  28.426  < 2e-16 ***
atmy                           -1.844e-02  7.770e-04 -23.728  < 2e-16 ***
hiv                            -4.094e-01  1.679e-02 -24.381  < 2e-16 ***
dpria                           3.460e-02  4.188e-03   8.261 2.25e-16 ***
bmi                             3.565e-02  4.673e-03   7.629 3.27e-14 ***
cAs                             2.862e+00  1.970e-01  14.527  < 2e-16 ***
cAm                             3.249e+00  2.178e-01  14.915  < 2e-16 ***
sts0                           -3.633e+00  2.493e-01 -14.577  < 2e-16 ***
gdp                             5.132e-05  6.464e-06   7.939 2.96e-15 ***
log(newdata2$under.five.deaths) -2.801e-01  4.490e-02  -6.239 5.11e-10 ***
plio                            2.157e-02  4.207e-03   5.127 3.15e-07 ***
th119                          -9.838e-02  2.189e-02  -4.495 7.26e-06 ***
cOc                             1.331e+00  3.520e-01   3.782 0.000159 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 13.20218)

    Null deviance: 223628  on 2696  degrees of freedom
Residual deviance:  35421  on 2683  degrees of freedom
AIC: 14629

Number of Fisher Scoring iterations: 2
```