World Health Organization Life Expectancy

Mark Vazquez

April 22, 2020

Introduction:

In the past, there have been many studies that show the factors that affect an individual's life expectancy. In these studies researchers have considered demographic variables, income composition and mortality rates. While conducting these studies researcher found that the effect of immunization and human development was not taken into account. Using one year's worth of data for all countries, some research was done considering multiple linear regression. The data from the Global Health Organization (GHO) under World Health Organization (WHO) contains data from all countries health status.

Purpose

Taking into account both of the affects that where described above we have found enough reason to formulate a regression model by using WHO data. We will focus on mixed effects model and multiple linear regression. We will also use data from the time period of 2000 to 2015 for all countries to model our regression. These models help countries determine where to invest more resources if they seek to improve overall Life Expectancy.

Multiple Linear Regression Overview

- y = Life Expectancy in age
- x1 = Adult Mortality
- x2 = Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
- x3 = Measles - number of reported cases per 1000 population
- x4 = Number of under-five deaths per 1000 population
- x5 = Polio (Pol3) immunization coverage among 1-year-olds (
- x6 = General Government Expenditure on health as a percentage of total government expenditure (
- x7 = Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (
- x8 = HIV/AIDSDeaths per 1 000 live births HIV/AIDS (0-4 years)
- x9 = GDPGross Domestic Product per capita (in USD)
- x10 = Population of the country
- x11 = thinness 1-19 yearsPrevalence of thinness among children
- x12 = Number of years of Schooling

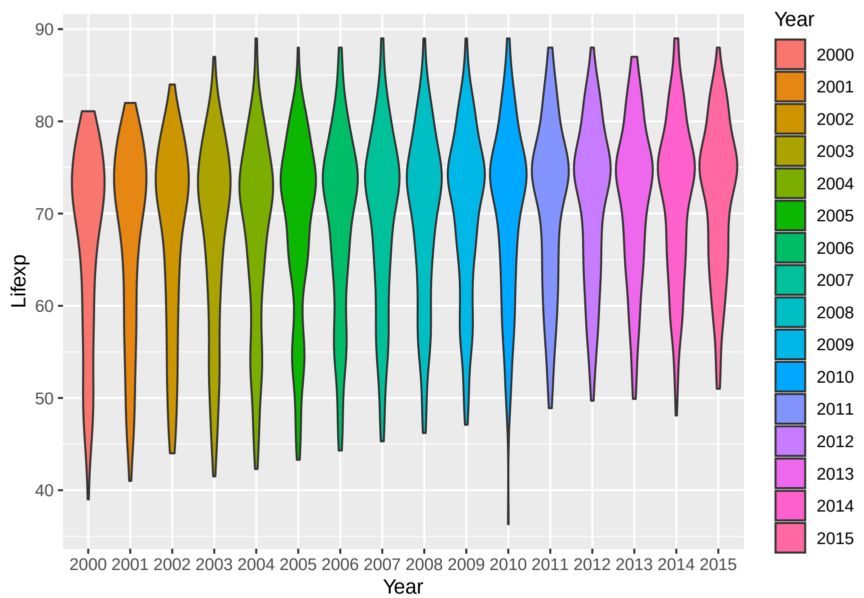
Data

Let us take a look at randomly selected rows

% latex table generated in R 4.0.0 by xtable 1.8-4 package % Sun May 03 19:31:37 2020

	Country	Year	Lifexp	Admort	Alcohol	Measles	U5deaths	Polio	Totexp	Diphtheria	HIV	GDP	Population	Thinness	Schooling
342	Botswana	2010	61.10	349	5.99	853	3	96	5.64	95	6.20	6434.82	2014866.00	8.00	12.30
1802	Namibia	2009	62.40	36	7.99	4076	4	83	8.50	83	8.70	5112.64	2137040.00	1.90	11.40
347	Botswana	2005	51.70	566	6.37	5	3	96	5.62	96	20.60	5686.78	1855852.00	1.00	11.90
177	Bahrain	2015	76.90	69		0		98		98	0.10	22435.73	1371855.00	6.20	14.50
1605	Maldives	2012	77.60	65	0.01	0		99	9.16	99	0.10	7251.68	386203.00	13.80	12.10
1504	Liberia	2001	51.50	333	4.40	1379	20	54	6.41	42	3.10	609.68	2991132.00	9.00	10.50
2223	Sao Tome and Principe	2008	65.40	215	4.36	0		99	5.66	99	0.90	1056.10	166913.00	6.50	10.40
589	Colombia	2003	72.40	15	4.25	0	19	92	5.92	92	0.10	5026.24	42152151.00	2.50	11.60
496	Cameroon	2000	51.40	394	3.91	14629	100	57	4.48	62	7.70	1145.45	15274234.00	7.70	6.90
1321	Japan	2008	82.70	66	7.11	11015	4	98	8.60	98	0.10	45165.79	128063000.00	1.80	15.00
1956	Pakistan	2000	62.80	19	0.02	2064	495	65	2.79	62	0.10	824.73	138523285.00	22.20	5.30
2567	Tajikistan	2000	63.70	198	0.37	192	17	86	4.64	83	0.30	415.46	6216205.00	4.20	9.60
1731	Mongolia	2000	62.80	274	2.79	925	3	94	4.92	94	0.10	1600.49	2397436.00	2.60	8.90
1432	Lao People's Democratic Republic	2009	63.10	223	5.18	78	14	67	3.77	67	0.20	1068.17	6152036.00	9.40	9.40
288	Benin	2000	55.40	279	1.34	4244	40	78	4.34	78	2.00	694.92	6865951.00	9.70	6.40

Data Visualization with Years(2000-2015)



Fitting Model

```r
summary(who.lm)
```

```
##
## Call:
## lm(formula = who_dat$Lifexp ~ . - Country - Year, data = who_dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.0800  -2.2788   0.0658   2.5038  12.2611
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.986e+01  6.364e-01  78.346  < 2e-16 ***
## Admort      -1.453e-02  9.369e-04 -15.513  < 2e-16 ***
## Alcohol     -1.573e-01  3.269e-02  -4.811 1.63e-06 ***
## Measles     -3.772e-05  9.161e-06  -4.117 4.02e-05 ***
```

```
## U5deaths    -3.214e-03  7.555e-04  -4.254 2.21e-05 ***
## Polio        9.985e-03  5.417e-03   1.843   0.0654 .
## Totexp       1.765e-02  4.413e-02   0.400   0.6893
## Diphtheria   3.830e-02  5.692e-03   6.729 2.29e-11 ***
## HIV         -5.524e-01  2.180e-02 -25.345  < 2e-16 ***
## GDP          9.468e-05  1.008e-05   9.396  < 2e-16 ***
## Population   7.156e-09  8.912e-10   8.030 1.75e-15 ***
## Thinness    -5.691e-02  2.523e-02  -2.256   0.0242 *
## Schooling    1.518e+00  5.008e-02  30.311  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.929 on 1791 degrees of freedom
##   (1076 observations deleted due to missingness)
## Multiple R-squared:  0.8338, Adjusted R-squared:  0.8327
## F-statistic: 748.6 on 12 and 1791 DF,  p-value: < 2.2e-16
```

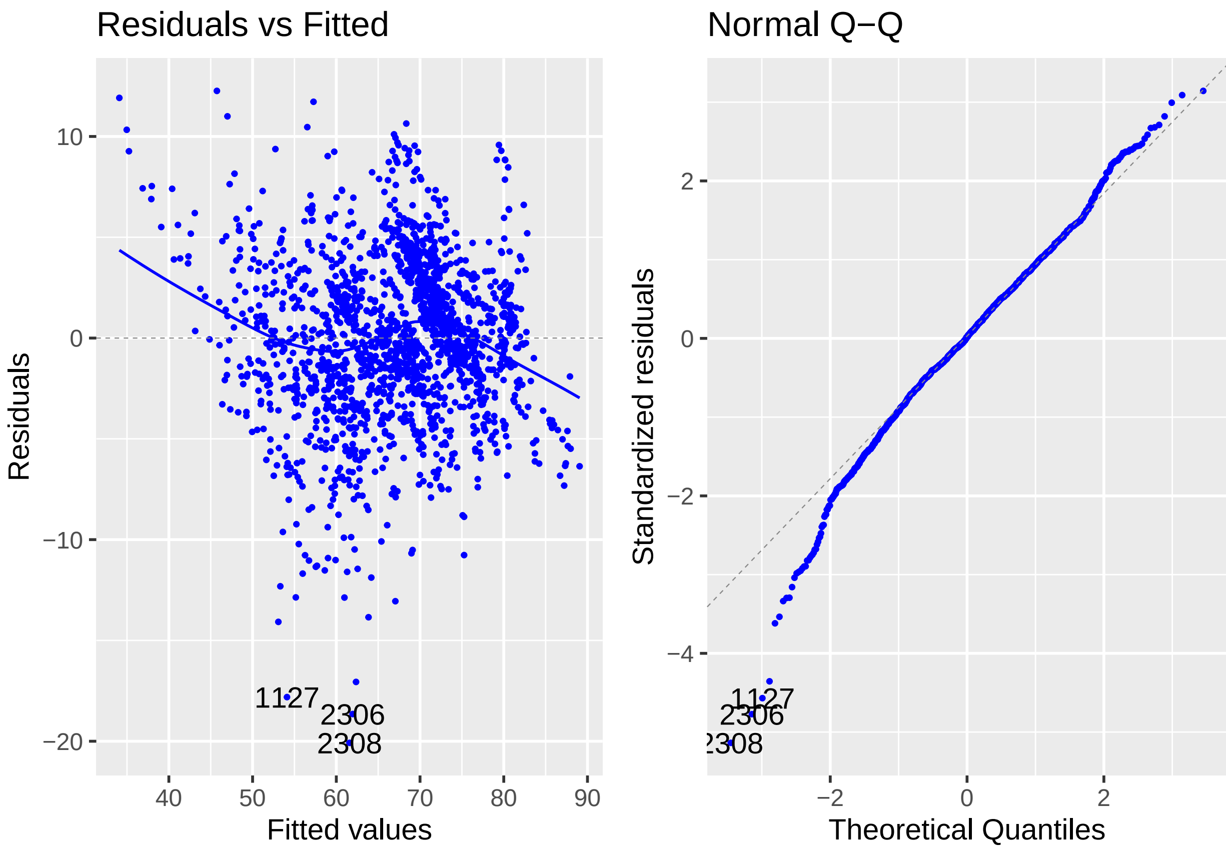
Our p-value: 2.2e-16 is significant at $\alpha = .05$. Therefore, we conclude that the model is significant. Hence, there is a linear relationship between the response y and any of the other of the regressor variables.

According to our t-tests the p values .0654,.6893 are greater than our significance level of $\alpha = .05$. Therefore the regressors Polio, Totexp are not contributing significantly to the model.

$$\hat{y} = 49.86 - .01453x_1 - .1573x_2 - .00003772x_3 - .003214x_4 + .00935x_5 + .01765x_6 + .03830x_7 - .5524x_8 + 00009468x_9$$

$$+ .000000007156x_{10} - .05691x_{11} + 1.518x_{12}$$

Model Adequacy Checking



In our residuals vs $\hat{y}$ there is not obvious pattern. Therefore, we satisfy Linearity assumption. However, our probability plot of the residuals may show issues with normaility. Thus, we proceed with the normality test.

```
ols_test_normality(who.lm)
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9865         0.0000
## Kolmogorov-Smirnov        0.0382         0.0105
## Cramer-von Mises        124.3553         0.0000
## Anderson-Darling          3.3569         0.0000
## -----------------------------------------------
```

According to the Shapiro-Wilk,Kolmogorov-Smirnov,Cramer-von Mises and Anderson-Darling normality tests, because of our small p value, we reject the null and conclude that our residuals are not normal. This maybe be due to influential observations

We procced to find some influential observations and potential outliers

```
inflm.fit <- influence.measures(who.lm)
inflm_obs <- which(apply(inflm.fit$is.inf, 1, any))
length(inflm_obs)
```

```
## [1] 158
```

There are 158 influential observations.

```
##
##  Shapiro-Wilk normality test
##
## data:  who.lm2$residuals
## W = 0.79906, p-value < 2.2e-16
```

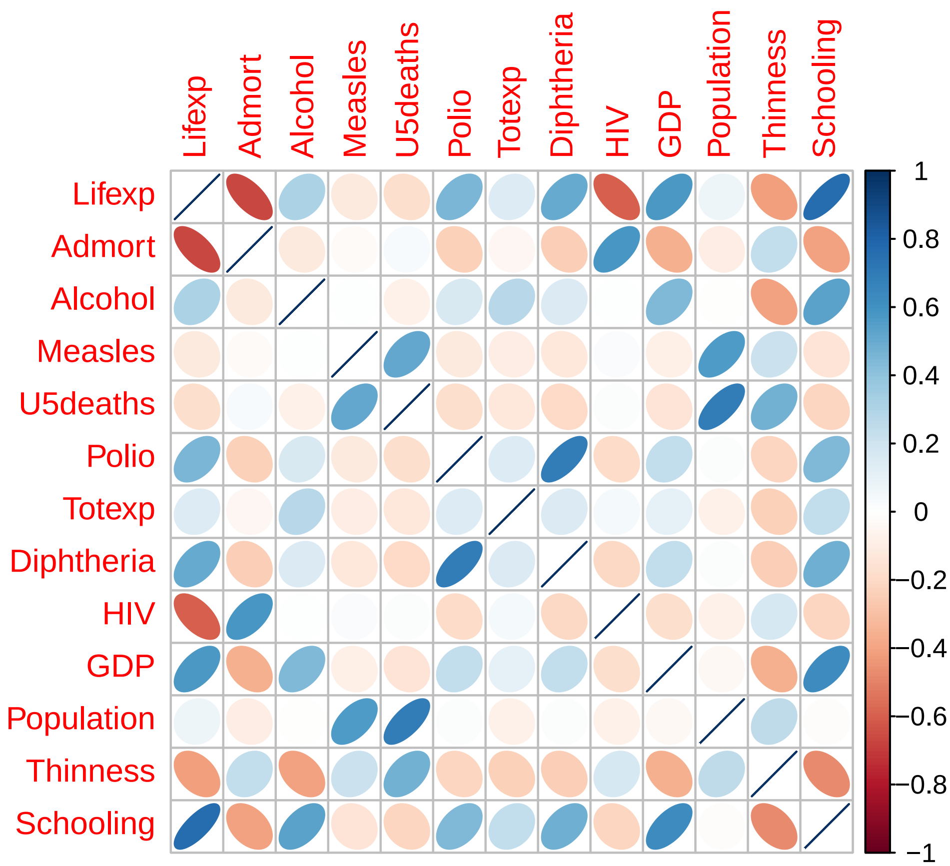
Removing influential observations with the largest residuals did not improve the adequacy of our model. Our normality assumptions were still not met. After some data exploratory analysis, these observations seem valid.

Examine correlation plot for any suspect of multicoliinearity

```
who_dat <- na.omit(who_dat) #Omit Na Values
who_dat <- select_if(who_dat, is.numeric) #Select only numeric Columns
row.names(who_dat) <- NULL #Resetting Index
#install.packages('corrplot')
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
corrplot(cor(who_dat), method = "ellipse")
```



There is suspect of multicollinearity because of some correlations between the variables

Thus we proceed to check Variance Inflation Factors

```
vif(who.lm)
```

```
##     Admort    Alcohol    Measles   U5deaths      Polio     Totexp Diphtheria
##   1.828499   1.679334   1.585609   2.592436   2.031997   1.141177   2.152803
```

```
##        HIV        GDP Population   Thinness  Schooling 
##   1.583458   1.781839   2.422343   1.758080   2.728819
```

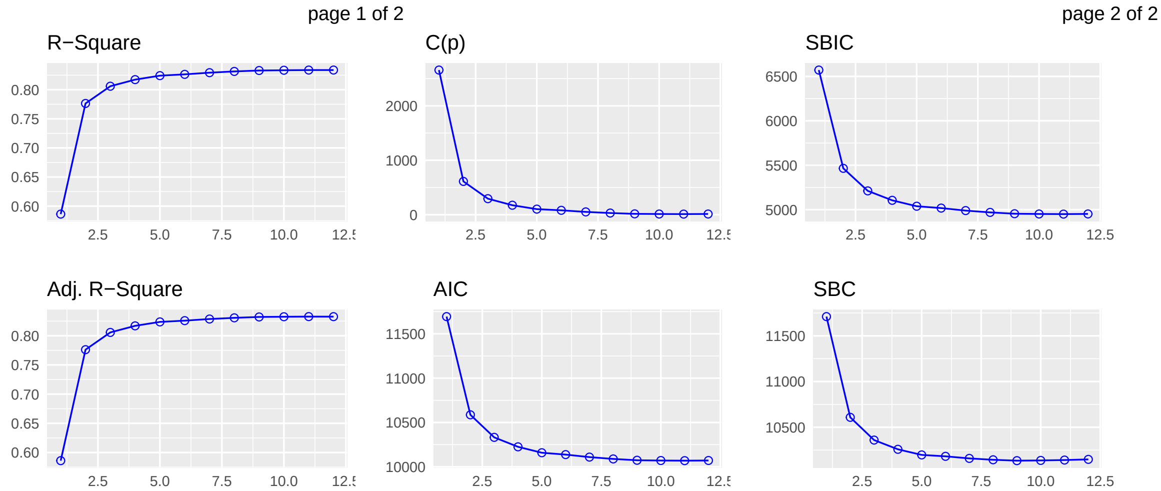
Our Variance Inflation Factors are all less than 5. We can conclude there is no multicollinearity issues.

Variable Selection and Model Building

```r
ols_step_best_subset(lm(Lifexp~., data=who_dat))
```

```
##                                            Best Subsets Regression                                            
## ------------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ------------------------------------------------------------------------------------------------------------
##      1         Schooling                                                                                      
##      2         HIV Schooling                                                                                  
##      3         Admort HIV Schooling                                                                           
##      4         Admort Diphtheria HIV Schooling                                                                
##      5         Admort Diphtheria HIV GDP Schooling                                                            
##      6         Admort Alcohol Diphtheria HIV GDP Schooling                                                    
##      7         Admort U5deaths Diphtheria HIV GDP Population Schooling                                        
##      8         Admort Alcohol U5deaths Diphtheria HIV GDP Population Schooling                                
##      9         Admort Alcohol Measles U5deaths Diphtheria HIV GDP Population Schooling                        
##     10         Admort Alcohol Measles U5deaths Diphtheria HIV GDP Population Thinness Schooling               
##     11         Admort Alcohol Measles U5deaths Polio Diphtheria HIV GDP Population Thinness Schooling         
##     12         Admort Alcohol Measles U5deaths Polio Totexp Diphtheria HIV GDP Population Thinness Schooling  
## ------------------------------------------------------------------------------------------------------------
## 
##                                                            Subsets Regression Summary                                                           
## ------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred                                                                    
## Model    R-Square    R-Square    R-Square      C(p)          AIC         SBIC          SBC          MSEP  
## ------------------------------------------------------------------------------------------------------------
##   1        0.5863      0.5861      0.5855    2656.8036    11693.9090    6571.2893    11710.4023    68863.3339  
##   2        0.7764      0.7761      0.7754     611.3661    10586.3261    5465.1620    10608.3171    37248.4796  
##   3        0.8060      0.8057      0.8045     294.3199    10332.0740    5211.4105    10359.5628    32334.0308  
##   4        0.8173      0.8169      0.8157     174.1288    10225.4121    5105.0341    10258.3987    30460.8548  
##   5        0.8242      0.8237      0.8224     102.5433    10158.6699    5038.5617    10197.1543    29338.2848  
##   6        0.8264      0.8258      0.8245      80.5823    10137.7085    5017.6775    10181.6906    28983.3608  
##   7        0.8293      0.8286      0.8272      51.5288    10109.5091    4989.6691    10158.9889    28518.0880  
##   8        0.8314      0.8306      0.8293      30.5897    10088.8716    4969.2165    10143.8492    28178.1633  
##   9        0.8330      0.8321      0.8308      15.5713    10073.8914    4954.4104    10134.3668    27929.7492  
##  10        0.8334      0.8325      0.8311      12.5915    10070.8925    4951.4775    10136.8656    27868.0040  
##  11        0.8338      0.8327      0.8312      11.1599    10069.4396    4950.0813    10140.9105    27830.2457  
##  12        0.8338      0.8327      0.8308      13.0000    10071.2785    4951.9369    10148.2472    27843.3071  
## ------------------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria 
##  SBIC: Sawa's Bayesian Information Criteria 
##  SBC: Schwarz Bayesian Criteria 
##  MSEP: Estimated error of prediction, assuming multivariate normality 
##  FPE: Final Prediction Error 
##  HSP: Hocking's Sp 
##  APC: Amemiya Prediction Criteria
```

```r
plot(ols_step_best_subset(lm(Lifexp~., data=who_dat)))
```



We will select the top 5 models and compare them

$$Model1 : \hat{y} = 49.86 + 1.518x_{12}$$

$$Model2 : \hat{y} = 49.86 + 1.518x_{12} + .5524x_8$$

$$Model3 : \hat{y} = 49.86 + 1.518x_{12} + .5524x_8 + .01453x_1$$

$$Model4 : \hat{y} = 49.86 + 1.518x_{12} + .5524x_8 + .01453x_1 + .03830x_7$$

$$Model5 : \hat{y} = 49.86 + 1.518x_{12} + .5524x_8 + .01453x_1 + .03830x_7 + 00009468x_9$$

PRESS Statistics for Models

```
## [1] "Our PRESS statistic for the First model is: 68925.9735616238"
## [1] "Our PRESS statistic for the Second model is: 37352.9349138681"
## [1] "Our PRESS statistic for the Third model is: 32501.8275243135"
## [1] "Our PRESS statistic for the Fourth model is: 30641.8984253618"
## [1] "Our PRESS statistic for the Fifth model is: 29526.8361755966"
```

Variance Inflation Factors for Models

```
## Schooling       HIV
##  1.048654  1.048654

## Schooling       HIV    Admort
##  1.195352  1.525211  1.737160

##  Schooling        HIV     Admort Diphtheria
##   1.477598   1.542060   1.737327   1.321474

##  Schooling        HIV     Admort Diphtheria        GDP
##   2.108629   1.544047   1.776499   1.331262   1.699836
```

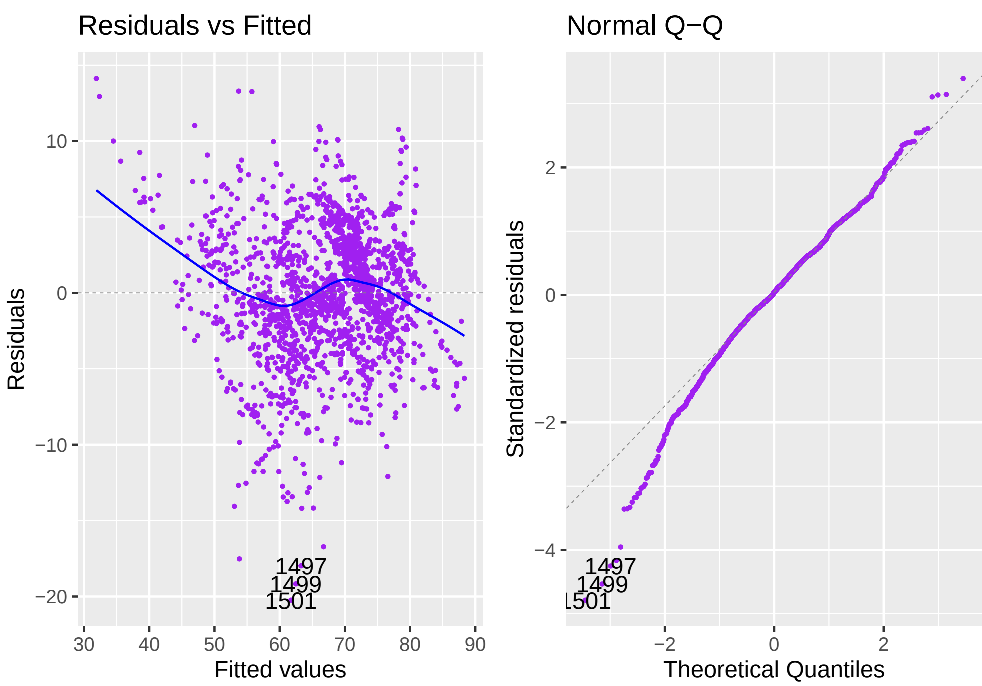
None of our models have issues with Multicollinearity. We proceed to select two models to compare. Parsimonious models are simple models with great explanatory predictive power. They explain data with a minimum number of parameters, or predictor variables. Therefore we picked models:

$$Model3 : \hat{y} = 49.86 + 1.518x_{12} + .5524x_8 + .01453x_1$$

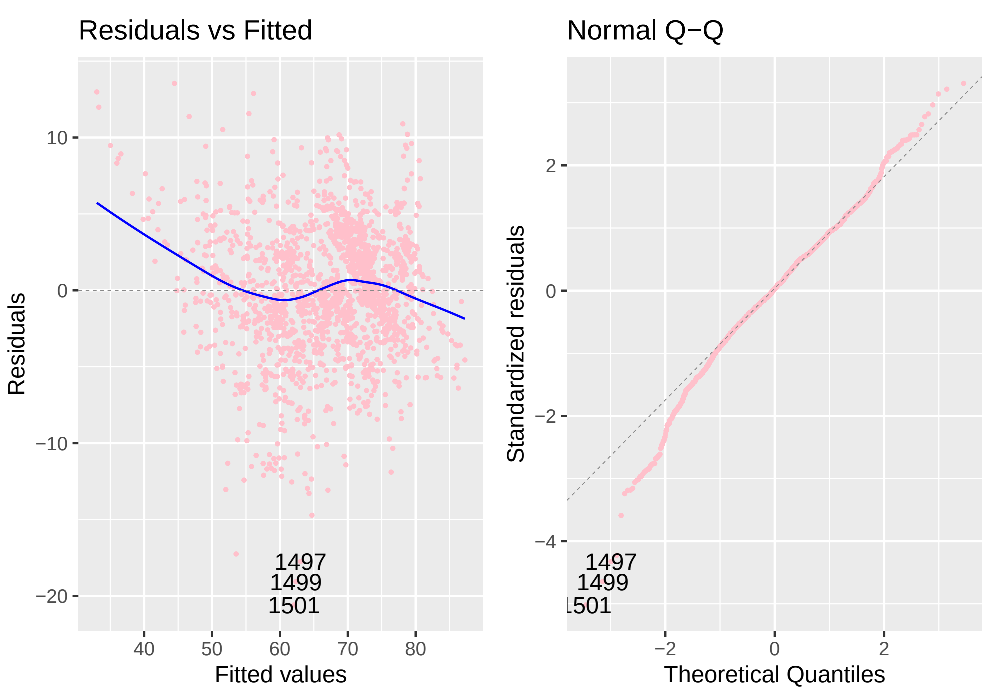
$$Model4: \hat{y} = 49.86 + 1.518x_{12} + .5524x_8 + .01453x_1 + .03830x_7$$

Due to their lower PRESS statistics compared to Models 1 and 2. They also have a high Pred R^2 values. .8045 and ,8157 respectively.

Model Adequacy of Model 3



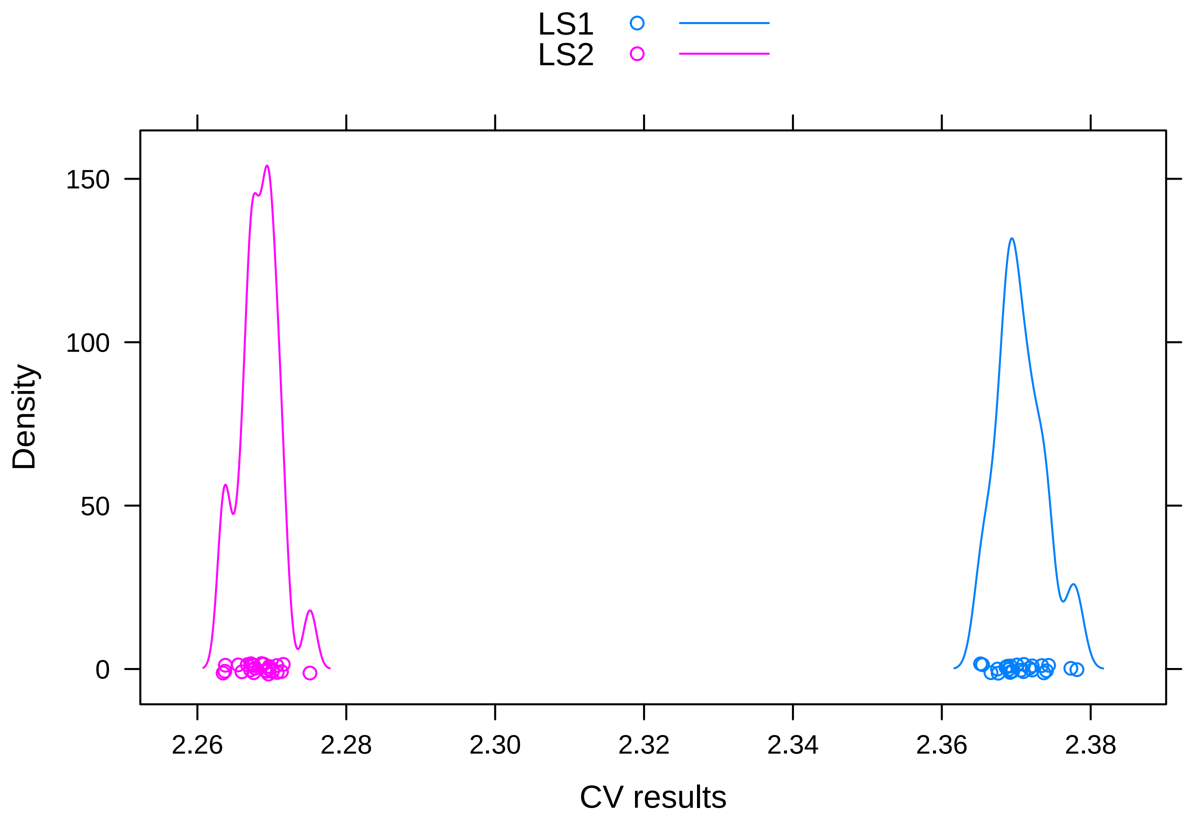
Model Adequacy of Model 4



For both our models, in our residuals vs $\hat{y}$ there is not obvious patterns. Therefore, we satisfy Linearity assumption.

Cross Validation

```
##
## 5-fold CV results:
##   Fit       CV
## 1 LS1 2.370626
## 2 LS2 2.268312
##
## Best model:
##    CV
## "LS2"
```




```r
summary(fit4)
```

```
##
## Call:
## lm(formula = Lifexp ~ Schooling + HIV + Admort + Diphtheria,
##     data = who_dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.585  -2.314   0.116   2.624  13.545
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 47.7536642  0.5216962   91.53   <2e-16 ***
## Schooling    1.7216144  0.0385466   44.66   <2e-16 ***
```

```
## HIV         -0.5641439  0.0224969  -25.08   <2e-16 ***
## Admort      -0.0164102  0.0009552  -17.18   <2e-16 ***
## Diphtheria   0.0492997  0.0046648   10.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.109 on 1799 degrees of freedom
## Multiple R-squared:  0.8173, Adjusted R-squared:  0.8169
## F-statistic:  2012 on 4 and 1799 DF,  p-value: < 2.2e-16
```

Conclusion:

We prefer $Model4 : \hat{y} = 49.86 + 1.518x_{12} + .5524x_8 + .01453x_1 + .03830x_7$ because it performed better in cross validation testing and because the principle of parsimony. The porportion of variation of the dependent variable (Life Expectancy) explained by this model is $R_{adj} = .8169$. It is interesting to note that the variables $x_{12} = Schooling, x_8 = HIV, x_1 = AdultMort, x_7 = Diptheria$ are significant variables in explaining Life Expectancy. Here, increasing Schooling by 1 unit(year) while holding all other variables constant, results an increase of 1.722 in Life Expectancy. We can also expect an imporvement in Life expectancy if we increase Diphtheria(immunization coverage) by one unit while holding other variables constant by 0.0493 years. Increasing HIV by 1 unit while holding all other constant results in a decrease of Life Expectancy(years) by -0.564. Similarly, a unit increase in Adult Mortalities while holding all other variables constant results in a decrease in Life Expectancy by -0.0164. These variables agree with our intuition. If countries provide education and immuzation, there is an overall increase in life expectancy. However, other variables like HIV and Adult mortalities negatively affect Life Expectancy. This is why HIV is a largely researched field and similarly overall adult health.

References

Data : https://www.kaggle.com/fahmadi96/life-expectancy-who-revised Montgomery, D., Peck, E., & Vining, G. (2012). Introduction to Linear Regression Analysis, 5th Edition. John Wiley & Sons.