ANALYSIS OF MULTICOLINEARITY AND NORMALITY ASSUMPTION OF SAMPLED DATA FROM NFHS 2005-06

**INTRODUCTION**

***Objective:*** Here our main objective is to detect the presence of multicolinearity and check the assumption of multicolinearity and we also want to check the normality assumption of the residuals of the fitted model of dataset sampled from NFHS 2005-06.

***Data Description:*** The dataset considered consist of 29 observations on 11 variables sampled from NFHS 2005-06. The dataset consist of following variables:

Y indicates TFR, dependent variable.

X1 indicates HDI.

X2 indicates Infant mortality rate.

X3 indicates contraceptive use (any method).

X4 indicates Female Age at marriage.

X5 indicates Median number of months since preceding the birth.

X6 indicates female literacy in percentage.

X7 indicates maternal care.

X8 indicates Male age at marriage.

X9 indicates percent of population with improved water supply.

X10 indicates male literacy in percentage.

Where x1,x2,…,x10 are all independent variables.

*#Reading NFHS dataset we are interested in.*  
**library**(readxl)  
NFHS <- **read\_excel**("NFHS.xlsx")  
  
*#Obtaining the first few records of the our dataset.*  
**head**(NFHS)

## # A tibble: 6 x 11  
## y x1 x2 x3 x4 x5 x6 x7 x8 x9 x10  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2.13 0.789 39.8 66.9 23.4 33.4 77.3 79.2 15.3 92.1 90.2  
## 2 2.69 0.644 41.7 63.4 41.4 30.4 60.4 42.1 27.7 95.6 83.4  
## 3 1.94 0.681 36.1 72.6 14.4 29.9 79.5 66 10.1 88.4 94   
## 4 2.38 0.601 44.7 52.6 16.1 32 53.9 77.2 14.4 80.8 78.1  
## 5 1.99 0.679 41.7 63.3 21.6 29.7 68.7 56.1 25.3 99.5 82.9  
## 6 3.21 0.537 65.3 47.2 58.4 30.2 36.2 33.9 49.2 81.8 73.9

**ANALYSIS**

*#Obtainig only the regressor variables from our dataset.*  
new\_NFHS=NFHS[2**:**10]  
**head**(new\_NFHS)

## # A tibble: 6 x 9  
## x1 x2 x3 x4 x5 x6 x7 x8 x9  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 0.789 39.8 66.9 23.4 33.4 77.3 79.2 15.3 92.1  
## 2 0.644 41.7 63.4 41.4 30.4 60.4 42.1 27.7 95.6  
## 3 0.681 36.1 72.6 14.4 29.9 79.5 66 10.1 88.4  
## 4 0.601 44.7 52.6 16.1 32 53.9 77.2 14.4 80.8  
## 5 0.679 41.7 63.3 21.6 29.7 68.7 56.1 25.3 99.5  
## 6 0.537 65.3 47.2 58.4 30.2 36.2 33.9 49.2 81.8

*#To obtain the coefficient of correlation matrix of regressors to check if there exist multicolinearity.*  
**round**(**cor**(new\_NFHS),2)

## x1 x2 x3 x4 x5 x6 x7 x8 x9  
## x1 1.00 -0.89 0.34 -0.74 0.35 0.89 0.65 -0.66 -0.07  
## x2 -0.89 1.00 -0.31 0.80 -0.36 -0.86 -0.76 0.77 0.03  
## x3 0.34 -0.31 1.00 -0.14 0.25 0.35 0.45 -0.34 0.51  
## x4 -0.74 0.80 -0.14 1.00 -0.26 -0.84 -0.60 0.77 0.19  
## x5 0.35 -0.36 0.25 -0.26 1.00 0.40 0.54 -0.53 -0.34  
## x6 0.89 -0.86 0.35 -0.84 0.40 1.00 0.64 -0.79 -0.09  
## x7 0.65 -0.76 0.45 -0.60 0.54 0.64 1.00 -0.82 -0.07  
## x8 -0.66 0.77 -0.34 0.77 -0.53 -0.79 -0.82 1.00 0.11  
## x9 -0.07 0.03 0.51 0.19 -0.34 -0.09 -0.07 0.11 1.00

***Inference:*** From the coefficient of correlation matrix we observe that there exist a highly negative correlation between x1 i.e. HDI and x2 i.e. Infant mortality rate,x1 i.e. HDI and x4 i.e. Female Age at marriage, x1 i.e. HDI and x8 i.e. Male age at marriage,x2 i.e. Infant mortality rate and x6 i.e. female literacy in percentage,x2 i.e. Infant mortality rate and x7 i.e. maternal care,x4 i.e. Female Age at marriage and x6 i.e. female literacy in percentage,x4 i.e. Female Age at marriage and x7 i.e. maternal care,x6 i.e. female literacy in percentage and x8 i.e. Male age at marriage,x7 i.e. maternal care and x8 i.e. Male age at marriage which means as the one variable increases the other decreases and highly positive correlation between x1 i.e. HDI and x6 i.e. female literacy in percentage,x7 i.e. maternal care and x1 i.e. HDI ,x2 i.e. Infant mortality rate and x4 i.e. Female Age at marriage,x2 i.e. Infant mortality rate and x8 i.e. Male age at marriage,x4 i.e. Female Age at marriage and x8 i.e. Male age at marriage,x6 i.e. female literacy in percentage and x7 i.e. maternal care which means as the one regressor variable increases the other also increases. Since there exist some correlation between the regressors it indicates the presence of multicolinearity.

*#Obtaining the fitted regression model to our dataset.*  
reg=**lm**(y**~**.,NFHS)  
reg

##   
## Call:  
## lm(formula = y ~ ., data = NFHS)  
##   
## Coefficients:  
## (Intercept) x1 x2 x3 x4 x5   
## 4.686737 0.613751 -0.001258 -0.033710 0.019574 -0.024824   
## x6 x7 x8 x9 x10   
## 0.005273 -0.017476 -0.007767 -0.006798 0.012862

*#Obtaining the summary of the fitted regression model.*  
**summary**(reg)

##   
## Call:  
## lm(formula = y ~ ., data = NFHS)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.43837 -0.14250 -0.04833 0.19676 0.35463   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.686737 1.970421 2.379 0.028660 \*   
## x1 0.613751 1.488450 0.412 0.684958   
## x2 -0.001258 0.009745 -0.129 0.898732   
## x3 -0.033710 0.007148 -4.716 0.000172 \*\*\*  
## x4 0.019574 0.008449 2.317 0.032514 \*   
## x5 -0.024824 0.024090 -1.030 0.316427   
## x6 0.005273 0.011015 0.479 0.637931   
## x7 -0.017476 0.006283 -2.782 0.012310 \*   
## x8 -0.007767 0.011157 -0.696 0.495198   
## x9 -0.006798 0.006318 -1.076 0.296125   
## x10 0.012862 0.018141 0.709 0.487409   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2751 on 18 degrees of freedom  
## Multiple R-squared: 0.8994, Adjusted R-squared: 0.8434   
## F-statistic: 16.08 on 10 and 18 DF, p-value: 5.185e-07

*#Loading the package 'car' required for obtaining the variance inflation factor.*  
**library**(car)

## Warning: package 'car' was built under R version 4.0.3

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.0.3

*#Obtaining the variance inflation factor for all the regressors of the fitted regression model.*  
**vif**(reg)

## x1 x2 x3 x4 x5 x6 x7 x8   
## 10.342818 8.857510 3.097699 6.585016 2.115003 11.024608 6.006303 6.787283   
## x9 x10   
## 2.254842 7.401606

***Inference:*** From the above table we observe that the resgressors x1, x2, x4, x6, x7, x8, x10 has VIF greater than 5 which indicates the critical level of multicolinearity hence we need to take the remedial measure to fix the multicolinearity we do it by observing that among all the regressor x6 i.e. female literacy in percentage has the highest variance inflation factor hence we try to fit the model by removing the regressor x6.

*#Now we are removing the regressor variable x6 from our dataset.*  
new1\_NFHS=NFHS[,**-**7]  
**head**(new1\_NFHS)

## # A tibble: 6 x 10  
## y x1 x2 x3 x4 x5 x7 x8 x9 x10  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2.13 0.789 39.8 66.9 23.4 33.4 79.2 15.3 92.1 90.2  
## 2 2.69 0.644 41.7 63.4 41.4 30.4 42.1 27.7 95.6 83.4  
## 3 1.94 0.681 36.1 72.6 14.4 29.9 66 10.1 88.4 94   
## 4 2.38 0.601 44.7 52.6 16.1 32 77.2 14.4 80.8 78.1  
## 5 1.99 0.679 41.7 63.3 21.6 29.7 56.1 25.3 99.5 82.9  
## 6 3.21 0.537 65.3 47.2 58.4 30.2 33.9 49.2 81.8 73.9

*#Fitting a new regression model to our dataset by removing the regressor x6 i.e female literacy from the dataset.*  
reg1=**lm**(y**~**.,new1\_NFHS)  
reg1

##   
## Call:  
## lm(formula = y ~ ., data = new1\_NFHS)  
##   
## Coefficients:  
## (Intercept) x1 x2 x3 x4 x5   
## 4.876626 0.989738 -0.001715 -0.032935 0.018210 -0.023376   
## x7 x8 x9 x10   
## -0.018585 -0.009947 -0.006975 0.013037

*#Obtaining the summary of fitted regression model of filtered dataset.*  
**summary**(reg1)

##   
## Call:  
## lm(formula = y ~ ., data = new1\_NFHS)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.42439 -0.16256 -0.06006 0.22185 0.32720   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.876626 1.890518 2.580 0.018370 \*   
## x1 0.989738 1.238412 0.799 0.434053   
## x2 -0.001715 0.009499 -0.181 0.858656   
## x3 -0.032935 0.006820 -4.829 0.000116 \*\*\*  
## x4 0.018210 0.007792 2.337 0.030529 \*   
## x5 -0.023376 0.023409 -0.999 0.330540   
## x7 -0.018585 0.005720 -3.249 0.004225 \*\*   
## x8 -0.009947 0.009976 -0.997 0.331253   
## x9 -0.006975 0.006178 -1.129 0.272942   
## x10 0.013037 0.017765 0.734 0.471996   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2695 on 19 degrees of freedom  
## Multiple R-squared: 0.8981, Adjusted R-squared: 0.8498   
## F-statistic: 18.6 on 9 and 19 DF, p-value: 1.277e-07

*#Obtaining the variance inflation factor of the regressors of new model.*  
**vif**(reg1)

## x1 x2 x3 x4 x5 x7 x8 x9   
## 7.462568 8.772490 2.938954 5.836637 2.081651 5.190271 5.656443 2.247148   
## x10   
## 7.398584

***Inference:*** From the above table we observe that the resgressors x1,x2,x4,x7,x8,x10 has VIF greater than 5 which again indicates the critical level of multicolinearity hence we need to take the remedial measure to fix the multicolinearity we do it by observing that among all the regressor x2 i.e. Infant mortality rate has the highest variance inflation factor hence we try to fit the model by removing the regressor x2.

*#Now we are removing the regressor variable x2 i.e. Infant mortality rate from our dataset.*  
new2\_NFHS=new1\_NFHS[,**-**3]  
**head**(new2\_NFHS)

## # A tibble: 6 x 9  
## y x1 x3 x4 x5 x7 x8 x9 x10  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2.13 0.789 66.9 23.4 33.4 79.2 15.3 92.1 90.2  
## 2 2.69 0.644 63.4 41.4 30.4 42.1 27.7 95.6 83.4  
## 3 1.94 0.681 72.6 14.4 29.9 66 10.1 88.4 94   
## 4 2.38 0.601 52.6 16.1 32 77.2 14.4 80.8 78.1  
## 5 1.99 0.679 63.3 21.6 29.7 56.1 25.3 99.5 82.9  
## 6 3.21 0.537 47.2 58.4 30.2 33.9 49.2 81.8 73.9

*#Fitting a new regression model to our dataset by removing the regressor x2 i.e. Infant mortality rate from the dataset.*  
reg2=**lm**(y**~**.,new2\_NFHS)  
reg2

##   
## Call:  
## lm(formula = y ~ ., data = new2\_NFHS)  
##   
## Coefficients:  
## (Intercept) x1 x3 x4 x5 x7   
## 4.673632 1.109315 -0.033280 0.017986 -0.023496 -0.018189   
## x8 x9 x10   
## -0.010187 -0.006674 0.013462

*#Obtaining the summary of fitted regression model of filtered dataset.*  
**summary**(reg2)

##   
## Call:  
## lm(formula = y ~ ., data = new2\_NFHS)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.40888 -0.16042 -0.05445 0.21929 0.33880   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.673632 1.482514 3.153 0.00501 \*\*   
## x1 1.109315 1.020747 1.087 0.29006   
## x3 -0.033280 0.006387 -5.211 4.24e-05 \*\*\*  
## x4 0.017986 0.007504 2.397 0.02642 \*   
## x5 -0.023496 0.022827 -1.029 0.31562   
## x7 -0.018189 0.005155 -3.528 0.00211 \*\*   
## x8 -0.010187 0.009646 -1.056 0.30352   
## x9 -0.006674 0.005803 -1.150 0.26366   
## x10 0.013462 0.017178 0.784 0.44243   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2629 on 20 degrees of freedom  
## Multiple R-squared: 0.8979, Adjusted R-squared: 0.8571   
## F-statistic: 21.99 on 8 and 20 DF, p-value: 2.63e-08

*#Obtaining the variance inflation factor of the regressors of new model.*  
**vif**(reg2)

## x1 x3 x4 x5 x7 x8 x9 x10   
## 5.327542 2.708743 5.688679 2.079978 4.429805 5.556510 2.083167 7.269030

***Inference:*** From the above table we observe that the resgressors x1,x4,x7,x8,x10 has VIF greater than 5 which indicates the critical level of multicolinearity and among all the regressor x10 i.e. male literacy in percentage has the highest variance inflation factor hence we try to fit the model by removing the regressor x10.

*#Now we are removing the regressor variable x10 i.e. male literacy from our dataset.*  
new3\_NFHS=new2\_NFHS[,**-**9]  
**head**(new3\_NFHS)

## # A tibble: 6 x 8  
## y x1 x3 x4 x5 x7 x8 x9  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 2.13 0.789 66.9 23.4 33.4 79.2 15.3 92.1  
## 2 2.69 0.644 63.4 41.4 30.4 42.1 27.7 95.6  
## 3 1.94 0.681 72.6 14.4 29.9 66 10.1 88.4  
## 4 2.38 0.601 52.6 16.1 32 77.2 14.4 80.8  
## 5 1.99 0.679 63.3 21.6 29.7 56.1 25.3 99.5  
## 6 3.21 0.537 47.2 58.4 30.2 33.9 49.2 81.8

*#Fitting a new regression model to our dataset by removing the regressor x10 i.e. male literacy from the dataset.*  
reg3=**lm**(y**~**.,new3\_NFHS)  
reg3

##   
## Call:  
## lm(formula = y ~ ., data = new3\_NFHS)  
##   
## Coefficients:  
## (Intercept) x1 x3 x4 x5 x7   
## 5.498365 1.661680 -0.031147 0.014956 -0.023664 -0.019678   
## x8 x9   
## -0.010649 -0.006536

*#Obtaining the summary of fitted regression model of filtered dataset.*  
**summary**(reg3)

##   
## Call:  
## lm(formula = y ~ ., data = new3\_NFHS)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.42812 -0.18480 -0.05302 0.22187 0.40237   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.498365 1.034519 5.315 2.86e-05 \*\*\*  
## x1 1.661680 0.731494 2.272 0.03374 \*   
## x3 -0.031147 0.005725 -5.441 2.13e-05 \*\*\*  
## x4 0.014956 0.006372 2.347 0.02879 \*   
## x5 -0.023664 0.022615 -1.046 0.30728   
## x7 -0.019678 0.004748 -4.144 0.00046 \*\*\*  
## x8 -0.010649 0.009539 -1.116 0.27686   
## x9 -0.006536 0.005746 -1.137 0.26822   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2605 on 21 degrees of freedom  
## Multiple R-squared: 0.8948, Adjusted R-squared: 0.8597   
## F-statistic: 25.51 on 7 and 21 DF, p-value: 6.612e-09

*#Obtaining the variance inflation factor of the regressors of new model.*  
**vif**(reg3)

## x1 x3 x4 x5 x7 x8 x9   
## 2.787192 2.216903 4.178324 2.079794 3.828383 5.535703 2.081239

From the above table we observe that the all the regressors have variance inflation factor between 1 to 5 which indicates moderate correlation between regressors.

Now we proceed to check for the normality assumption of residuals of fitted model.

*#Obtaining the residuals of the final fitted model.*  
res=**resid**(reg3)  
  
*#Performing shapiro wilks normality test to check for normality assumption.*  
**shapiro.test**(res)

##   
## Shapiro-Wilk normality test  
##   
## data: res  
## W = 0.9599, p-value = 0.3268

***Inference:*** Since the p value = 0.3268 is greater than 0.05 we accept the null hypothesis and conclude that the normality assumptions are satisfied and we say that errors are normally distributed.