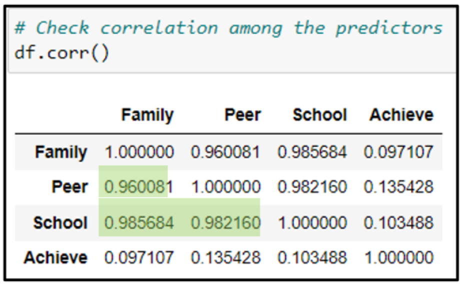
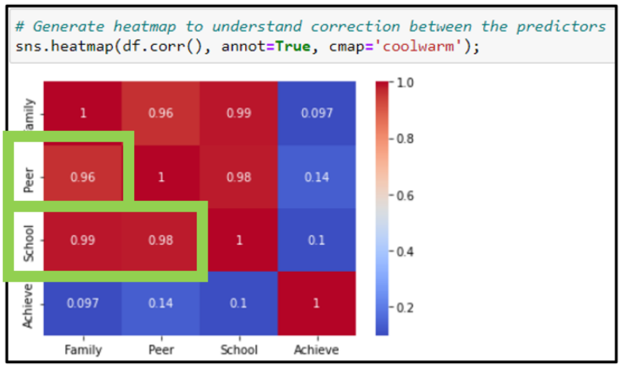
1) Data Exploration

The excel file was read. The quantitative response variable (‘Achieve’) and the predictor variables (‘Family’, ‘Peer’ and ‘School’) were being explored. In particular, the below correlation table and heatmap were generated by executing the python scripts.



```
# Check correlation among the predictors
df.corr()
```

	Family	Peer	School	Achieve
Family	1.000000	0.960081	0.985684	0.097107
Peer	0.960081	1.000000	0.982160	0.135428
School	0.985684	0.982160	1.000000	0.103488
Achieve	0.097107	0.135428	0.103488	1.000000



```
# Generate heatmap to understand correction between the predictors
sns.heatmap(df.corr(), annot=True, cmap='coolwarm');
```

	Family	Peer	School	Achieve
amily	1	0.96	0.99	0.097
Peer	0.96	1	0.98	0.14
School	0.99	0.98	1	0.1
Achieve	0.097	0.14	0.1	1

Diagram 1: Correlation Table and Heatmap generated on the variables to be studied

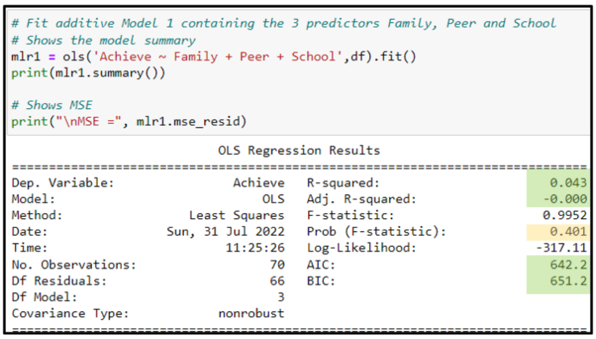
** All python scrips used are presented on top of the results

It was observed from both the correlation table (values shaded in green) and heatmap (values boxed in green) that the ‘Family’, ‘Peer’ and ‘School’ predictor variables are strongly correlated.

2) Fitting of 3 models

(i) Model 1: Additive MLR model

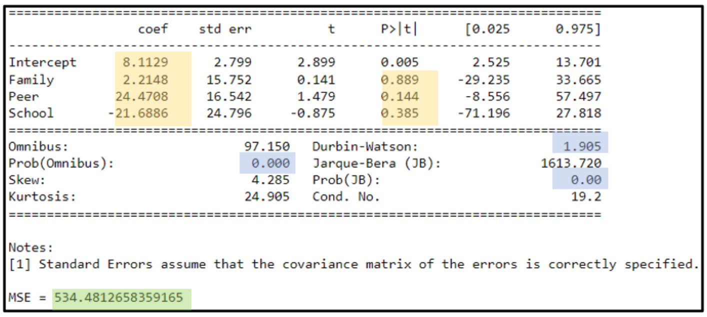
The below additive MLR model was fitted.



```
# Fit additive Model 1 containing the 3 predictors Family, Peer and School
# Shows the model summary
mlr1 = ols('Achieve ~ Family + Peer + School',df).fit()
print(mlr1.summary())

# Shows MSE
print("\nMSE =", mlr1.mse_resid)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Achieve   R-squared:                       0.043
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.9952
Date:                Sun, 31 Jul 2022   Prob (F-statistic):              0.401
Time:                        11:25:26   Log-Likelihood:                -317.11
No. Observations:                  70   AIC:                             642.2
Df Residuals:                      66   BIC:                             651.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
```



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.1129      2.799      2.899      0.005       2.525      13.701
Family         2.2148     15.752      0.141      0.889     -29.235      33.665
Peer          24.4708     16.542      1.479      0.144      -8.556      57.497
School       -21.6886     24.796     -0.875      0.385     -71.196      27.818
==============================================================================
Omnibus:                       97.150   Durbin-Watson:                   1.905
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1613.720
Skew:                           4.285   Prob(JB):                         0.00
Kurtosis:                      24.905   Cond. No.                         19.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

MSE = 534.4812658359165
```

Diagram 2: Summary and MSE details for MLR Model 1 fitted

** Values used for goodness of fit and accuracy are shaded in green, values on test statistics and coefficient are shaded in yellow, and values from tests conducted which used in model assumptions are shaded in blue. These shaded values were summarised in Table 1 comparison of models, as well as referenced in this document.

The MLR Model1 equation can be expressed as:

$$\widehat{\text{Achieve}} = 8.1129 + 2.2148 * \text{Family} + 24.4708 * \text{Peer} - 21.6886 * \text{School}$$

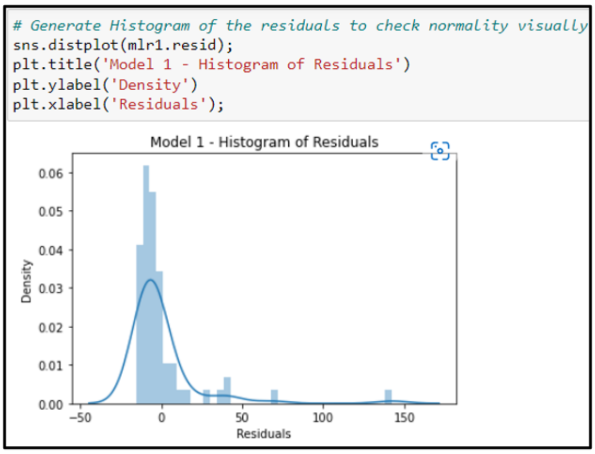
After fitted the Model 1, it's residuals were studied to verify if the assumptions needed for the linear regression model were met. It was observed that the following 2 assumptions are NOT complied with:

- Residuals should follow a normal distribution
- There should be little or no multicollinearity among the predictors

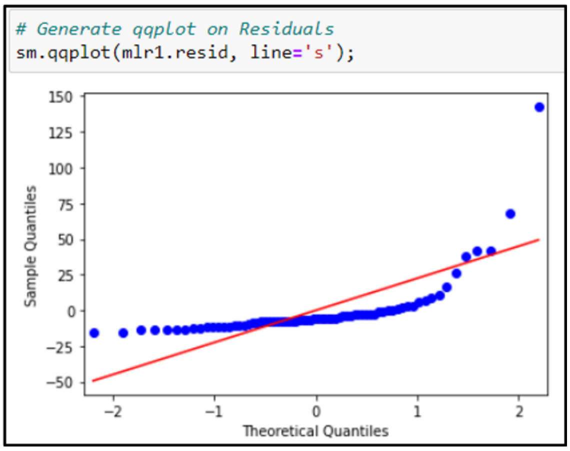
(ii) Model 2: Investigation of log-transformation for response variable 'Achieve'

On assumption where the residuals of model 1 not following a normal distribution, the following verification were carried out:

- Plotted the histogram on residuals, where it should show a bell-shape distribution. The generated distribution is skewed to the right as shown in Diagram 3
- Plotted the normal probability plot (or QQ plot), where it's points should fall roughly on a straight line. The points in the generated plot does not fall on the straight line, as shown in Diagram 3
- Conducted Omnibus and Jarque-Bera normality tests, indicating that sufficient evidence at 5% of significance to conclude that data are not normally distributed.

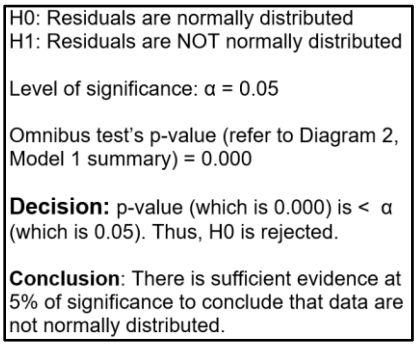


```
# Generate Histogram of the residuals to check normality visually
sns.distplot(mlr1.resid);
plt.title('Model 1 - Histogram of Residuals')
plt.ylabel('Density')
plt.xlabel('Residuals');
```



```
# Generate qqplot on Residuals
sm.qqplot(mlr1.resid, line='s');
```

Diagram 3: Histogram on residuals showing right skewed distribution (left), and QQ plot showing data points not falling on the red line (right), for model 1



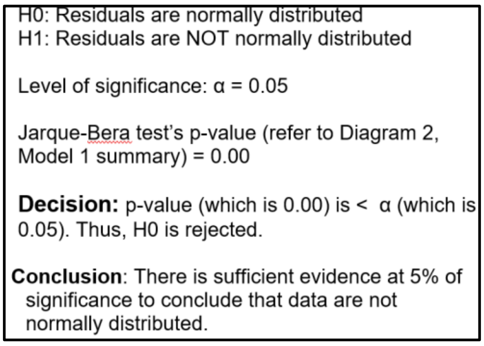
H0: Residuals are normally distributed
H1: Residuals are NOT normally distributed

Level of significance: α = 0.05

Omnibus test's p-value (refer to Diagram 2, Model 1 summary) = 0.000

Decision: p-value (which is 0.000) is < α (which is 0.05). Thus, H0 is rejected.

Conclusion: There is sufficient evidence at 5% of significance to conclude that data are not normally distributed.



H0: Residuals are normally distributed
H1: Residuals are NOT normally distributed

Level of significance: α = 0.05

Jarque-Bera test's p-value (refer to Diagram 2, Model 1 summary) = 0.00

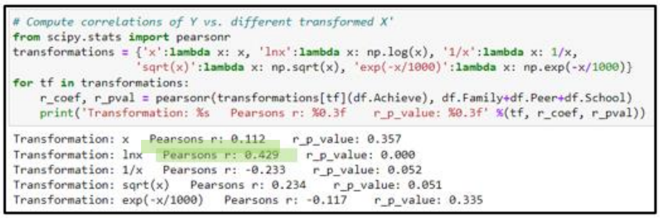
Decision: p-value (which is 0.00) is < α (which is 0.05). Thus, H0 is rejected.

Conclusion: There is sufficient evidence at 5% of significance to conclude that data are not normally distributed.

Diagram 4: Omnibus Normality Test (left) and Jarque-Bera Normality tests (right) for model 1

To improve the normality distribution, usually the response variable is to be transformed.

The following python script to compute the correlation of response variable vs different transformed values is executed. The results showed that the log (lnx) transformation generates the closest score to 1.

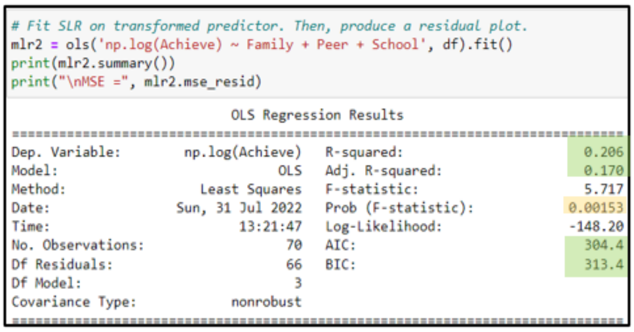


```
# Compute correlations of Y vs. different transformed X'
from scipy.stats import pearsonr
transformations = {'x':lambda x: x, 'lnx':lambda x: np.log(x), '1/x':lambda x: 1/x,
                   'sqrt(x)':lambda x: np.sqrt(x), 'exp(-x/1000)':lambda x: np.exp(-x/1000)}
for tf in transformations:
    r_coef, r_pval = pearsonr(transformations[tf](df.Achieve), df.Family+df.Peer+df.School)
    print('Transformation: %s   Pearsons r: %0.3f    r_p_value: %0.3f' %(tf, r_coef, r_pval))

Transformation: x   Pearsons r: 0.112    r_p_value: 0.357
Transformation: lnx   Pearsons r: 0.429    r_p_value: 0.000
Transformation: 1/x   Pearsons r: -0.233    r_p_value: 0.052
Transformation: sqrt(x)   Pearsons r: 0.234    r_p_value: 0.051
Transformation: exp(-x/1000)   Pearsons r: -0.117    r_p_value: 0.335
```

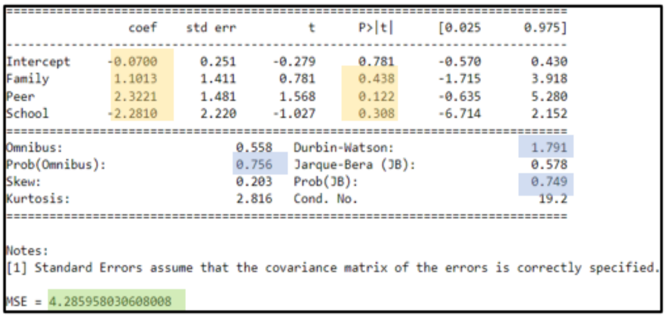
Diagram 5: Compute different Pearson r scores for correlation between the response and the transformed variables.

Thus, this log transformation on the response variable was chosen and model 2 was fitted.



```
# Fit SLR on transformed predictor. Then, produce a residual plot.
mlr2 = ols('np.log(Achieve) ~ Family + Peer + School', df).fit()
print(mlr2.summary())
print("\nMSE =", mlr2.mse_resid)

                            OLS Regression Results
==============================================================================
Dep. Variable:        np.log(Achieve)   R-squared:                       0.206
Model:                            OLS   Adj. R-squared:                  0.170
Method:                 Least Squares   F-statistic:                     5.717
Date:                Sun, 31 Jul 2022   Prob (F-statistic):            0.00153
Time:                        13:21:47   Log-Likelihood:                -148.20
No. Observations:                  70   AIC:                             304.4
Df Residuals:                      66   BIC:                             313.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
```



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     -0.0700      0.251     -0.279      0.781      -0.570       0.430
Family         1.1013      1.411      0.781      0.438      -1.715       3.918
Peer           2.3221      1.481      1.568      0.122      -0.635       5.280
School        -2.2810      2.220     -1.027      0.308      -6.714       2.152
==============================================================================
Omnibus:                        0.558   Durbin-Watson:                   1.791
Prob(Omnibus):                  0.756   Jarque-Bera (JB):                0.578
Skew:                           0.203   Prob(JB):                        0.749
Kurtosis:                       2.816   Cond. No.                         19.2
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

MSE = 4.285958030608008
```

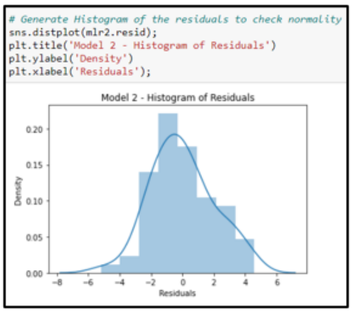
Diagram 6: Summary and MSE details for MLR Model 2 using log-transformation of response variable is fitted

The MLR Model 2 equation can be expressed as:

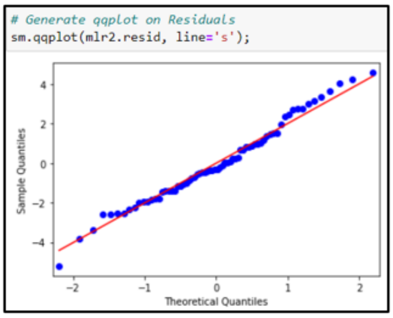
$$log(\widehat{\text{Achieve}}) = -0.07 + 1.1013 * \text{Family} + 2.3221 * \text{Peer} - 2.2810 * \text{School}$$

Next, the histogram and QQ plot were generated for Model 2 to verify resolution of the residual normality. In the below diagram, both plots demonstrated normal distribution after the log linear transformation.

Based on the model summary, both p-values for Omnibus and Jarque-Bera normality tests (=0.756 and 0.749 respectively) are > α (= 0.05). Thus, there is insufficient evidence at 5% of significance to conclude that the data are not normally distributed.



```
# Generate Histogram of the residuals to check normality
sns.distplot(mlr2.resid);
plt.title('Model 2 - Histogram of Residuals')
plt.ylabel('Density')
plt.xlabel('Residuals');
```



```
# Generate qqplot on Residuals
sm.qqplot(mlr2.resid, line='s');
```

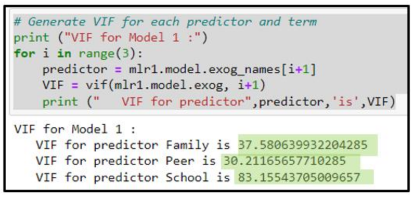
Diagram 7: Histogram on residuals showing normal bell shape (left), and QQ plot showing data points falling on the red line (right), for model 2

(iii) Model 3: Investigate multicollinearity and perform Principal Component Regression

The multicollinearity assumption for linear regression model is also NOT complied with.

The followings indicate the presence of data-based multicollinearity:

- High (strong) correlations between the predictors 'Family', 'Peer' and 'School', as indicated in the correlation matrix and heatmap in the Diagram 1.
- Predictors 'Family', 'Peer' and 'School' have high Variance Inflation Factor (VIF) of 37, 30 and 83 respectively, which are much higher than 5 (=highly correlated).



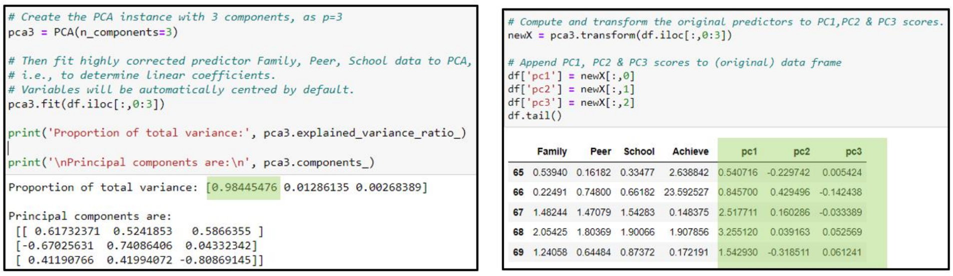
```
# Generate VIF for each predictor and term
print ("VIF for Model 1 :")
for i in range(3):
    predictor = mlr1.model.exog_names[i+1]
    VIF = vif(mlr1.model.exog, i+1)
    print ("   VIF for predictor",predictor,'is',VIF)
```

```
VIF for Model 1 :
   VIF for predictor Family is 37.580639932204285
   VIF for predictor Peer is 30.21165657710285
   VIF for predictor School is 83.15543705009657
```

Diagram 8: VIF for Model 1

To resolve the multicollinearity problem, Principal Component Analysis (PCA) can be used to find the "best" linear combinations for the 3 highly corrected predictors, followed by adapting the desired PC (replacing the highly correlated predictors) in the model.

The following python script is executed to find the PCA, compute the PC values and then added the PC values to the data frame.



```
# Create the PCA instance with 3 components, as p=3
pca3 = PCA(n_components=3)

# Then fit highly corrected predictor Family, Peer, School data to PCA,
# i.e., to determine linear coefficients.
# Variables will be automatically centred by default.
pca3.fit(df.iloc[:,0:3])

print('Proportion of total variance:', pca3.explained_variance_ratio_)

print('\nPrincipal components are:\n', pca3.components_)
```

```
Proportion of total variance: [0.98445476 0.01286135 0.00268389]

Principal components are:
 [[ 0.61732371  0.5241853   0.5866355 ]
 [-0.67025631  0.74086406  0.04332342]
 [ 0.41190766  0.41994072 -0.80869145]]
```

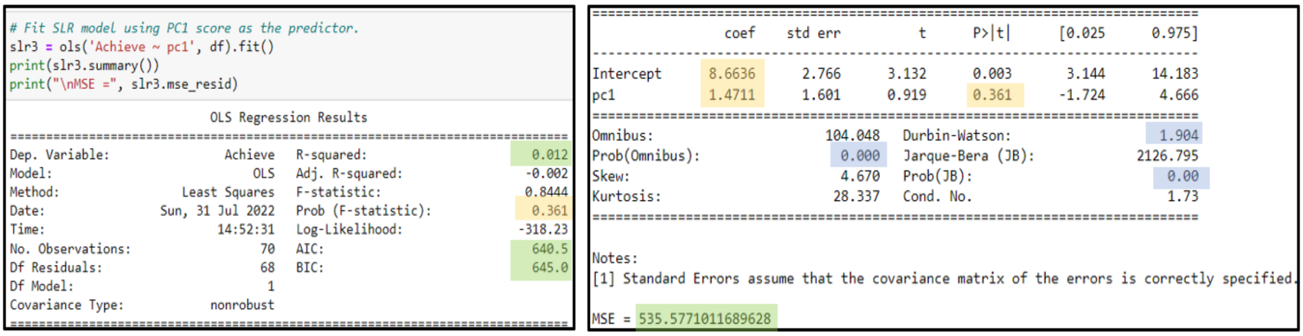
```
# Compute and transform the original predictors to PC1,PC2 & PC3 scores.
newX = pca3.transform(df.iloc[:,0:3])

# Append PC1, PC2 & PC3 scores to (original) data frame
df['pc1'] = newX[:,0]
df['pc2'] = newX[:,1]
df['pc3'] = newX[:,2]
df.tail()
```

	Family	Peer	School	Achieve	pc1	pc2	pc3
65	0.53940	0.16182	0.33477	2.638842	0.540716	-0.229742	0.005424
66	0.22491	0.74800	0.66182	23.592527	0.845700	0.429496	-0.142438
67	1.48244	1.47079	1.54283	0.148375	2.517711	0.160286	-0.033389
68	2.05425	1.80369	1.90066	1.907856	3.255120	0.039163	0.052569
69	1.24058	0.64484	0.87372	0.172191	1.542930	-0.318511	0.061241

Diagram 9: Performed PCA, transformed the linear combinations into PC and added them to the data frame.

PC1 accounts for 98.4% of the total variability, and thus Model 3 was fitted using just the PC1 as follows:



```
# Fit SLR model using PC1 score as the predictor.
slr3 = ols('Achieve ~ pc1', df).fit()
print(slr3.summary())
print("\nMSE =", slr3.mse_resid)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Achieve   R-squared:                       0.012
Model:                            OLS   Adj. R-squared:                 -0.002
Method:                 Least Squares   F-statistic:                    0.8444
Date:                Sun, 31 Jul 2022   Prob (F-statistic):              0.361
Time:                        14:52:31   Log-Likelihood:                -318.23
No. Observations:                  70   AIC:                             640.5
Df Residuals:                      68   BIC:                             645.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      8.6636      2.766      3.132      0.003       3.144      14.183
pc1            1.4711      1.601      0.919      0.361      -1.724       4.666
==============================================================================
Omnibus:                      104.048   Durbin-Watson:                   1.904
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2126.795
Skew:                           4.670   Prob(JB):                         0.00
Kurtosis:                      28.337   Cond. No.                         1.73
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

MSE = 535.5771011689628
```

Diagram 10: Summary and MSE details for SLR Model 3 using PC1

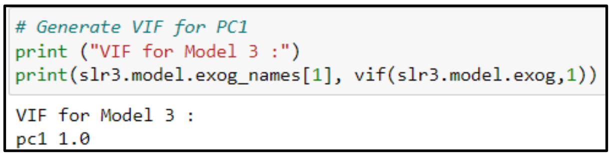
The MLR Model 3 equation can be expressed as:

$\widehat{\text{Achieve}} = 8.6636 + 1.4711 * \text{PC1}$, where

PC1 = 0.61732371*(X_{Family} $X_{\text{Family Center}}$) + 0.5241853*(X_{Peer} - $X_{\text{Peercenter}}$) + 0.5866355*(X_{School} - $X_{\text{School Center}}$)

PC1 = 0.61732371*(X_{Family} - 0.049383) + 0.5241853*(X_{Peer} - 0.046314) + 0.5866355*(X_{School} – 0.031906)

The VIF for Model 3 is now 1, verifying that that the multicollinearity is resolved.



```
# Generate VIF for PC1
print ("VIF for Model 3 :")
print(slr3.model.exog_names[1], vif(slr3.model.exog,1))
```

```
VIF for Model 3 :
pc1 1.0
```

Diagram 11: VIF for Model 3

3) Comparison of the 3 Models

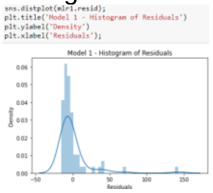
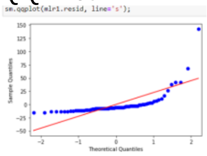
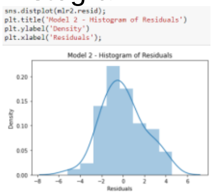
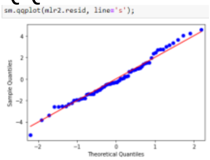
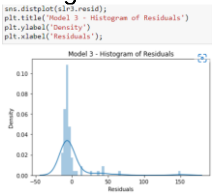
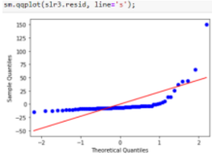
Refer to below table 1 for details of comparison.

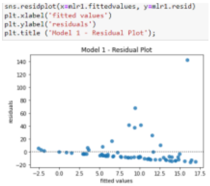
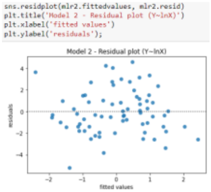
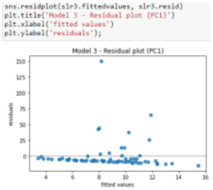
In summary:

- All the 3 models do not provide good model fit as the coefficient of determination is low, and the test statistics indicate that the predictors are statistically insignificant
- Of the 3 models, Model 3 might be the better choice due to its more favourable AIC and BIC values, and with absence of multicollinearity

	Metrics	Model 1	Model 2	Model 3	Explanation
(a) Goodness of fit	(i) Coefficient of determination, R^2 or adjusted R^2	R^2 = 0.043 Adjusted R^2 = -0.000	R^2 = 0.206 Adjusted R^2=0.170	R^2 = 0.012	• For SLR, R^2 is more useful. For MLR, adjusted R^2 is more useful, as a penalty factor is applied to insignificant predictors. • All the 3 models have low values, indicating that the predictors do not account much proportion of the total variation in the response variable.
	(ii) AIC, BIC	• AIC = 642.2 • BIC = 651.2	• AIC = 304.4 • BIC = 313.4	• AIC = 640.5 • BIC = 645.0	• The lower AIC and BIC, the better is the model fit. • Model 3 is a better model fit than Model 1. • It is not meaningful to compare Model 2 against Model 1 and 3, as it is of different scaling.
(b) Accuracy for prediction	(i) MSE	534.481	4.286	535.577	• Mean Square Error (MSE) measures the prediction accuracy of model. The lower the MSE, the higher is the prediction accuracy. • Thus, model 1 is slightly more accurate for prediction than model 3
(c) Statistical Tests for prediction	(i) F-test on overall model	0.401	0.002	0.361	As the p-value for Model 1 and 3 (=0.401 and 0.361 respectively) is > α (=0.05), do not reject H0. There is insufficient evidence at 5% of significance to conclude that model contains at least one significant predictor.
	(ii) t-tests on coefficients	p-values are 0.889 (Family), 0.144 (Peer), 0.385 (School)	p-values are 0.438 (Family), 0.122 (Peer), 0.308 (School)	p-value for PC1 is 0.361	In all 3 models, each predictors' p-values are > α (=0.05), thereby H0 is not rejected. Thus, for all the predictors in the 3 models, there is insufficient evidence at 5% of significance to conclude that each predictor is statistically significant.

	Metrics	Model 1	Model 2	Model 3	Explanation

(d) Meeting assumption for linear regression model	(i) Residuals follows a normal distribution
	- Via plotting histogram on residuals
	- Via QQ plot
	- Conduct Omnibus test
	- Conduct Jarque-Bera (JB) test
		Histogram:
		
		`sns.distplot(mlr1.resid);`
		`plt.title('Model 1 - Histogram of Residuals')`
		`plt.ylabel('Density')`
		`plt.xlabel('Residuals');`
		
		QQ Plot:
		
		`sm.qqplot(mlr1.resid, line='s');`
		
		Prob(Omnibus) = 0.000
		Prob(JB)= 0.00
			Histogram:
			
			`sns.distplot(mlr2.resid);`
			`plt.title('Model 2 - Histogram of Residuals')`
			`plt.ylabel('Density')`
			`plt.xlabel('Residuals');`
			
			QQ Plot:
			
			`sm.qqplot(mlr2.resid, line='s');`
			
			Prob(Omnibus) = 0.756
			Prob(JB)= 0.749
				Histogram:
				
				`sns.distplot(slr3.resid);`
				`plt.title('Model 3 - Histogram of Residuals')`
				`plt.ylabel('Density')`
				`plt.xlabel('Residuals');`
				
				QQ Plot:
				
				`sm.qqplot(slr3.resid, line='s');`
				
				Prob(Omnibus) = 0.000
				Prob(JB)= 0.00
					<u>For Model 1 and 3:</u>
					Both the residual histogram and QQplot shows it is skewed to the right.
					The omnibus and Jarque-Bera tests both have p-values < α (=0.05). Thus, H0 is rejected. There is sufficient evidence at 5% of significance to conclude that data are not normally distributed.
					<u>For Model 2:</u>
					The residuals follow normal distribution.

	(ii) Residuals have constant variance (homoscedastic)
	- Via plotting residuals vs fitted value
	- Conduct Breusch Pagan test
		Plot:
		
		`sns.residplot(x=mlr1.fittedvalues, y=mlr1.resid)`
		`plt.xlabel('fitted values')`
		`plt.ylabel('residuals')`
		`plt.title ('Model 1 - Residual Plot');`
		
		P-value for Breusch Pagan 0.302
			Plot:
			
			`sns.residplot(mlr2.fittedvalues, mlr2.resid)`
			`plt.title('Model 2 - Residual plot (Y~lnX)')`
			`plt.xlabel('fitted values')`
			`plt.ylabel('residuals');`
			 Model 2 - Residual plot (Y~lnX)
			residuals
			fitted values
			P-value for Breusch Pagan 0.944
				Plot:
				
				`sns.residplot(slr3.fittedvalues, slr3.resid)`
				`plt.title('Model 3 - Residual plot (PC1)')`
				`plt.xlabel('fitted values')`
				`plt.ylabel('residuals');`
				
				P-value for Breusch Pagan 0.345
					Model 1, 2 and 3 have Breusch Pagan p-values (=0.302, 0.944 and 0.345 respectively) which are > α (=0.05).
					Thus, do not reject H0.
					There is insufficient evidence at 5% of significance to conclude that the residuals are heteroscedastic.

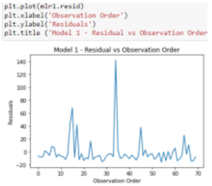
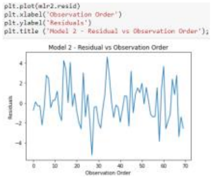
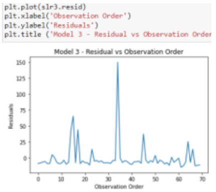
	Metrics	Model 1	Model 2	Model 3	Explanation
	(iii) Residuals are independent - Plot residuals vs observed order - Conduct Durbin-Watson Test	Plot:  plt.plot(mlr1.resid) plt.xlabel('Observation Order') plt.ylabel('Residuals') plt.title ('Model 1 - Residual vs Observation Order Model 1 - Residual vs Observation Order Residuals Observation Order 140 120 100 80 60 40 20 0 −20 0 10 20 30 40 50 60 70 Durbin-Watson 1.905	Plot:  plt.plot(mlr2.resid) plt.xlabel('Observation Order') plt.ylabel('Residuals') plt.title ('Model 2 - Residual vs Observation Order'); Model 2 - Residual vs Observation Order Residuals Observation Order 4 2 0 −2 −4 0 10 20 30 40 50 60 70 Durbin-Watson 1.791	Plot:  plt.plot(slr3.resid) plt.xlabel('Observation Order') plt.ylabel('Residuals') plt.title ('Model 3 - Residual vs Observation Order Model 3 - Residual vs Observation Order Residuals Observation Order 150 125 100 75 50 25 0 0 10 20 30 40 50 60 70 Durbin-Watson 1.904	With the plot showing random pattern, and the Durbin-Watson value close to 2, there is no autocorrelation for all 3 models.
	(iv) Little or no multicollinearity among the predictors - Observe Correlation matrix - Derive High Variance Inflation Factor (VIF)	VIF for predictors: • Family is 37.581 • Peer is 30.212 • School is 83.155	VIF for predictors: • Family is 37.581 • Peer is 30.211 • School is 83.155	VIF for PC1 is 1	The predictors in Model 1 and Model 2 have high VIF (VIF is high if > 5). Thus, Model 1 and Model 2 have multicollinearity problem. There is no multicollinearity problem with Model 3.

Table 1: Comparison of the 3 models fitted

** All figures in above table are rounded to 3 decimal places.

<< End of Report >>