I. Proposal

The Problem, the Variables, and the Data Collection Process

In this research project, our focus is on assessing the factors influencing college and career readiness rates in New York City high schools. We will be utilizing the "2017 DOE High School Directory" dataset, which provides comprehensive information about various aspects of high schools in the city. The primary objective of our study is to build a predictive model capable of estimating college and career readiness rates based on selected high school characteristics.

Our analysis is anchored by the 'College Career Rate' variable, which serves as our dependent variable. This measure represents the percentage of high school students deemed ready for college and career based on predetermined criteria. The following four predictor variables, ‘attendance_rate’, ‘graduation_rate’, ‘pct_stu_enough_variety’, and ‘pct_stu_safe’, will be investigated to understand their impact on college and career readiness:

Data Collection Process:

The dataset was sourced from the official data repository of the City of New York, specifically the "2017 DOE High School Directory" available at the following website.
https://data.cityofnewyork.us/Education/2017-DOE-High-School-Directory/s3k6-pzi2.

Meaning for Modeling this Dataset:

Understanding the factors that contribute to college and career readiness is crucial for educators, policymakers, and students. Our research endeavor aims to shed light on the intricate relationship between high school characteristics and the preparedness of students for higher education and career pursuits. By modeling this dataset, we aspire to offer insights into the following areas:

1. **Educational Policy:** Policymakers can benefit from a better understanding of how graduation rates, attendance, course variety, and students' perceived safety impact college and career readiness. This knowledge can inform strategic decisions to enhance educational outcomes.

2. **School Improvement Strategies:** Educators and school administrators can use the predictive model to identify areas of improvement related to graduation rates, attendance, course offerings, and safety, ultimately enhancing the overall preparedness of students for future endeavors.

Scatter Plots of the Response Variable vs. Each Predictor Variable

Response Variable:

College_Career_Rate (Continuous): The 'College Career Rate' represents the percentage of high school students deemed ready for college and career pursuits. It is our primary response variable, and our goal is to predict college_career_rate based on the following four predictor variables.

Predictor Variables:

1. **Graduation_Rate (Continuous):** The percentage of students who complete their high school education. We aim to explore how graduation rates influence college and career readiness.

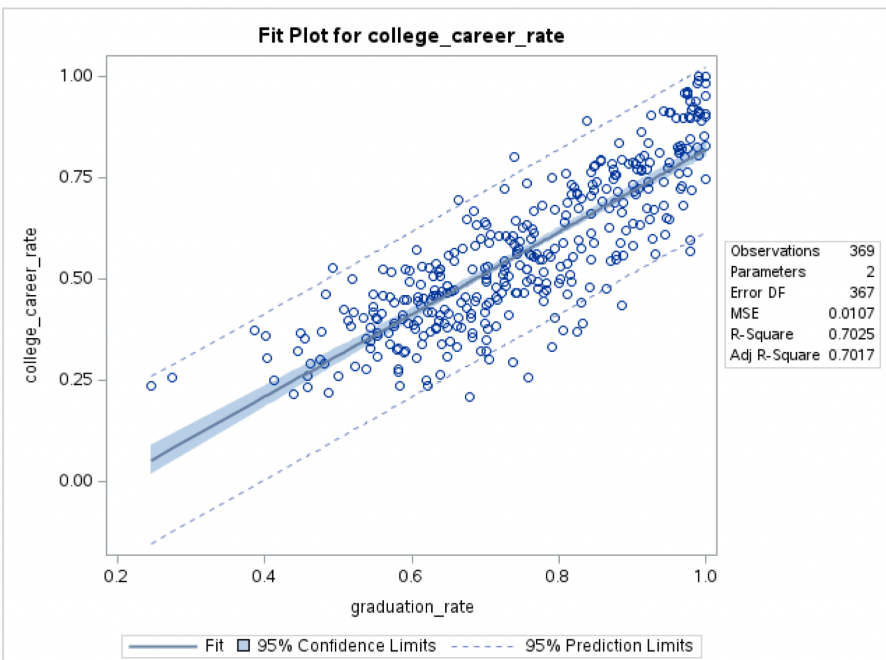
2. **Attendance_Rate (Continuous):** The percentage of student attendance during the academic year. We seek to understand how attendance rates correlate with college and career readiness.

3. **Pct_Stu_Enough_Variety (Continuous):** Represents the variety of courses available to students. We intend to explore how this factor relates to college and career readiness.

4. **Pct_Stu_Safe (Continuous):** Reflects the perceived safety of students within the school environment. We aim to investigate the connection between students' sense of safety and their readiness for college and careers.

In subsequent sections, we will visualize the relationships between 'College Career Rate' and each of the selected predictor variables through scatter plots and conduct regression analysis to gain a deeper understanding of their influence on college and career readiness.

Figure 1: college_career_rate Vs. graduation_rate



In Figure 1, we see an upward trend and a promising adjusted R-squared value of 0.7017. This pair has the highest r-squared value among all pairs. This figure has an upward trend with no curvature and could be our best plot fit for regression analysis.

Figure 2: college_career_rate Vs. attendance_rate

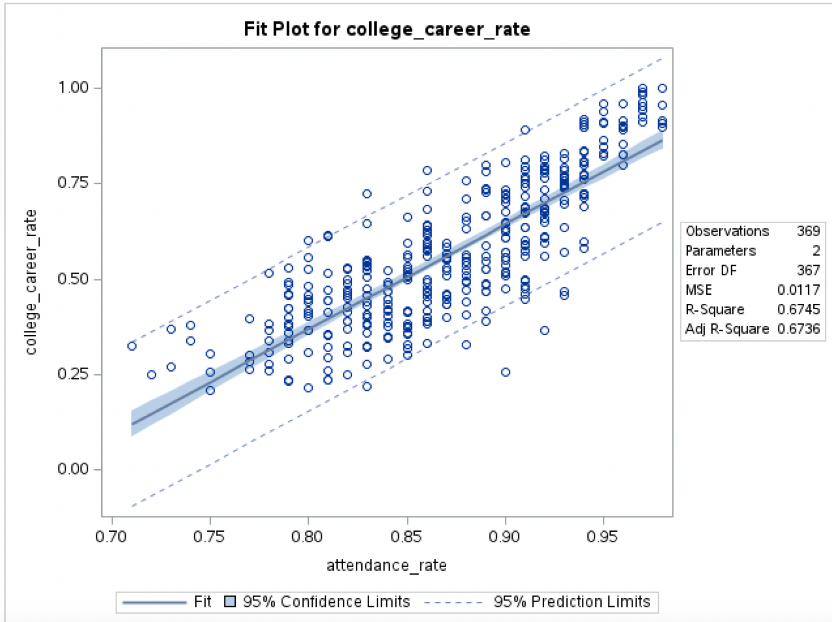


Figure 2 shows an adjusted R-squared value of 0.6736, but it may not be good enough for our regression analysis. Although we see a linear trend and a good R-squared value, graduation_rate seems to be a better predictor for the college_career_rate than the attendance_rate.

Figure 3: college_career_rate Vs. pct_stu_safe

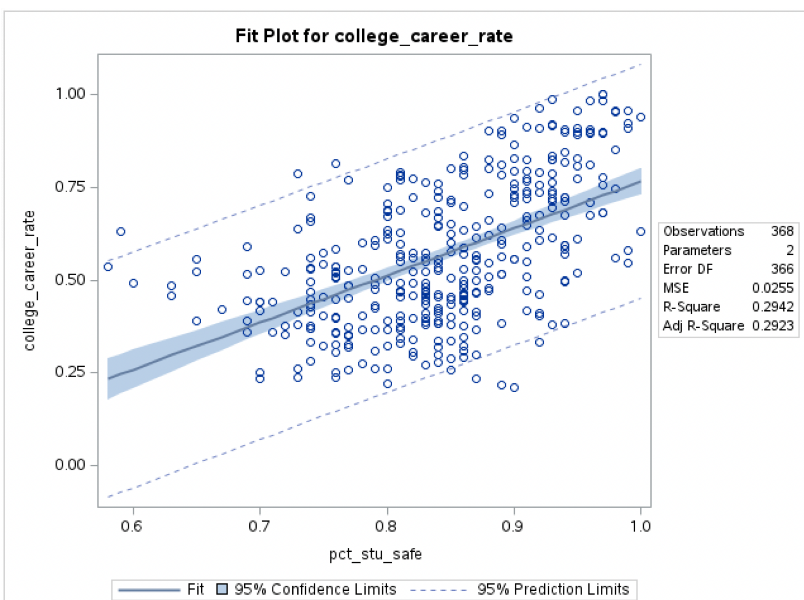


Figure 3 shows a slight linear trend. The adjusted R-squared value is 0.2923, which is low for linear regression analysis. This pair would be a bad choice for regression analysis, so we'll look at the other pairs.

Figure 4: college_career_rate Vs. pct_stu_enough_variety

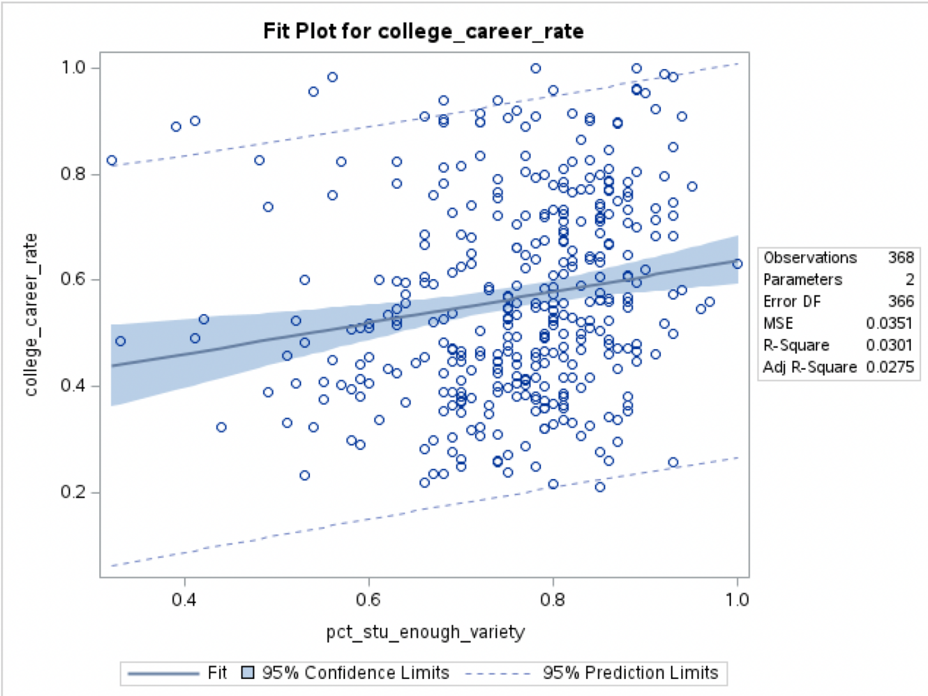


Figure 4 shows an even lower adjusted R-squared value of 0.0275. We don't see a good slope created by the data points and it looks like a random point cloud. Pct_stu_enough_variety is not a good choice as our predictor variable.

Based on our scatterplots, graduation_rate seems to be the best fit for a Simple Linear Regression analysis. Let's explore why this is the best pick in the following section.

Selection of the predictor variable:

Taking the scatter plot observations into consideration we have concluded that 'graduation_rate' will be the best predictor variable for the response variable 'college_career_rate'. We can observe from the Figure 1 scatterplot that this pair provides us with the best correlation value, which is 0.7017. It shows a good correlation between 'graduation_rate' and 'college_career_rate'. The scatterplot for this pair also indicates the presence of regression.

Other predictor variables studied will be taken out of consideration for the Simple Linear Regression due to weak correlation with our response variable and lack of evidence of the existence of regression.

II. Simple Linear Regression Model

Simple linear regression takes the form: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where, Y_i is the response variable, which is the college_career_rate, and Xi is the predictor variable, which is the graduation_rate.

β0 and β1 are the regression coefficients, where they represent the intercept and slope of the regression line respectively. εi is the random error term associated with the model.

1. **Fitting the regression line:**

Figure 5: college_career_rate vs. graduation_rate.

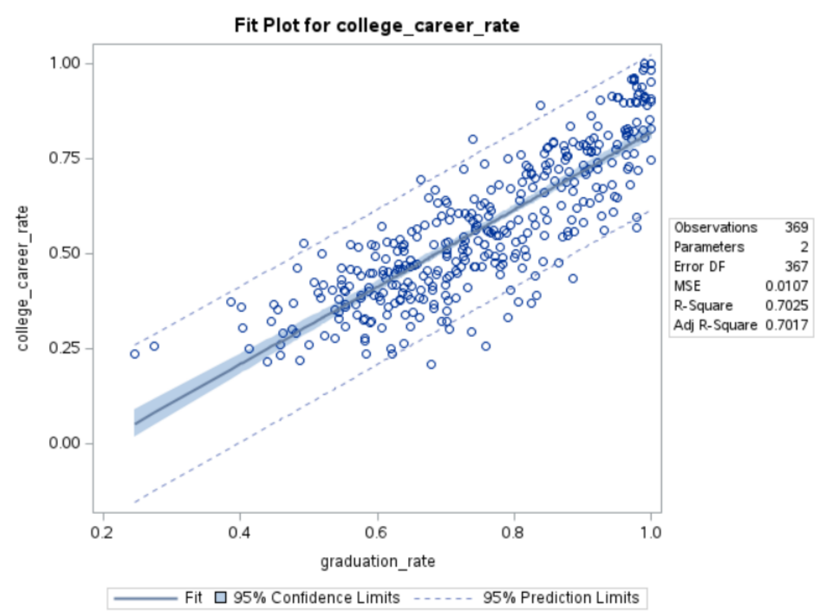


Figure 5 shows that there exists a linear relationship between college_career_rate and graduation_rate.

Table 1: ANOVA table for college_career_rate vs graduation_rate.

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Number of Observations Read	440
Number of Observations Used	369
Number of Observations with Missing Values	71

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	9.30390	9.30390	866.77	<.0001
Error	367	3.93938	0.01073		
Corrected Total	368	13.24328			

Root MSE	0.10361	R-Square	0.7025
Dependent Mean	0.56672	Adj R-Sq	0.7017
Coeff Var	18.28159		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.19928	0.02657	-7.50	<.0001
graduation_rate	1	1.01886	0.03461	29.44	<.0001

The point estimator for b0 is −0.19928 which represents the estimated value of college_career_rate when graduation_rate is equal to 0. The point estimator for b1 is 1.01886 indicating that, on average, for each additional unit of graduation_rate, college_career_rate is estimated to increase by 1.01886 units. Therefore, the simple linear regression model of our analysis is represented as:

$College\_\widehat{career}\_rate$ = **−0.19928+ 1.0188 graduation_rate**

The model obtained an R-Square value of 0.7025, which indicates a linear association between college_career_rate and graduation_rate.

2. **Inferences on the Parameters**

Confidence interval for the slope:

Hypothesis Test:
H_0: $\beta_1 = 0$
VS.
H_1: $\beta_1 \neq 0$
95% confidence interval for β_1(slope):
Here, α = 1 - 0.95 = 0.05

Two sided confidence interval:
$b_1 + t(1 - \frac{\alpha}{2}, n\text{-}2) \times s\{b_1\}$ ——-(1)

$s\{b1\} = \sqrt{\frac{MSE}{\Sigma x_i^2 - (\Sigma x_i)^2/n}}$

$\Sigma x_i^2 - (\Sigma x_i)^2/n = \sum(\ )^2 = (n\text{-}1)\times s_x^2 = (369\text{-}1)\times(0.1565468)^2 = 9.0185$

$s\{b1\} = \sqrt{\frac{0.01073}{9.0185}} = 0.03449$

From t-distribution table, $t(1 - \frac{\alpha}{2}, n\text{-}2) = t(0.975,367) = 1.96$

Substituting all the above values in equation

$\Rightarrow b_1 + t(1 - \frac{\alpha}{2}, n\text{-}2) \times s\{b_1\}$

$\Rightarrow 1.01886 \pm 1.96\times0.03449$

$\Rightarrow$ CI for β_1= (0.95126, 1.08646)

Conclusion:
We are 95% confident that the mean college_career_rate lies between 0.95126 and 1.08646
$\beta_1 = 0$ is not in the confidence interval, Hence we reject H_0 gives us that the linear regression is significant.

Confidence interval for the y-intercept:

Hypothesis Testing:
H_0: $\beta_0 = 0$
VS.
H_1: $\beta_0 \neq 0$
95% confidence interval for β_0(Intercept):
Here, α = 1 - 0.95 = 0.05
Two sided confidence interval:
$b_0 + t(1 - \frac{\alpha}{2}, n\text{-}2) \times S\{b_0\}$ — (2)

$$s\{b_0\} = \sqrt{MSE(\frac{1}{n} + \frac{\overline{(x)}^2}{\Sigma x_i^2 - (\Sigma x_i)^2/n})}$$

$$\Sigma x_i^2 - (\Sigma x_i)^2/n = \sum(\,)^2 = (n-1)\times s_x^2 = (369-1)\times(0.1565468)^2 = 9.0185$$

$$s\{b_0\} = \sqrt{0.01073(\frac{1}{369} + \frac{(0.7524413)^2}{9.0185})} = 0.02650$$

From t-distribution table, $t(1 - \frac{\alpha}{2}, n\text{-}2) = t(0.975,367) = 1.96$

Substituting all the above values in equation (2)

$\Rightarrow$ $b_0 + t(1 - \frac{\alpha}{2}, n\text{-}2) \times s\{b_0\}$
$\Rightarrow$ $-0.19928 \pm 1.96\times0.02650$

$\Rightarrow$CI for β_0= (-0.25122, -0.14734)

Conclusion:
We are 95% confident that the y-intercept is significant since $\beta_0 = 0$ does not lies in the confidence interval, we reject H_0. Even though the test shows the y-intercept is significant the intercept is not meaningful.

3. Confidence Interval and Prediction Interval for a new (x_h)

We are interested in a new graduation_rate (x_h) = 0.807 with alpha (α) = 0.05 .

$\hat{Y}_h = b_0 + b_1 \times x_h$

$= -0.19928 + 1.01886 \times 0.807$

$= 0.62294002$

Now, we can calculate the error for the predicted college_career_rate.

$$s\{\hat{Y}_h\} = \sqrt{MSE\left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\Sigma x_i^2 - (\Sigma x_i)^2/n}\right)}$$

$\Sigma x_i^2 - (\Sigma x_i)^2/n = \sum(\ )^2 = (n-1)s_x^2 = (369-1) \times 0.156850^2$

$= 368 \times 0.024601922$

$= 9.05350748$

$s\{\hat{Y}_h\} = \sqrt{0.01073[(1/369) + ((0.807 - 0.752802)^2/9.05350748)}$

$= \sqrt{0.01073(0.00271 + 0.00032445)} = 0.00570611$

The 95% confidence interval is given by:

$CI = \hat{Y}_h \pm t(1 - \frac{\alpha}{2}, n\text{-}2) \times s\{\hat{Y}_h\}$

$= 0.62294002 \pm 1.96 \times 0.00570611$

$= (0.5045432, 0.74133682)$

Conclusion: We are 95% confident that the mean college_career_rate for a school with graduation_rate(X_h) = 0.807 lies between 0.5045432 and 0.74133682.

Now, we will calculate the prediction interval to account for the additional uncertainty of predicting a new observation.

$s\{pred\} = \sqrt{(MSE + s\{\hat{Y}_h\}^2)} = \sqrt{0.01073 + 0.0000325546}$

$= 0.103742732$

Prediction Interval $= \hat{Y}_h \pm t(1 - \frac{\alpha}{2}, n\text{-}2) \times s\{pred\}$

$= 0.62294002 \pm 1.96 \times 0.103742732$

$= (0.41960426, 0.82627578)$

Conclusion: We are 95% confident that the next observation of college_career_rate for a school with graduation_rate(X_h) = 0.807 will lie between 0.41960426 and 0.82627578.

4. **Working-hotelling confidence bands for our regression function**

Xh (point to predict) = 0.46

To calculate the 95% confidence band, we first need to calculate the 95% critical value from the F-distribution with n-2 degrees of freedom, where n is the sample size used to estimate the regression model.

The F-critical value for 95% confidence with df1(numerator) = and df2(denominator)= 367 is equal to 3.

The formula for the 95% confidence band is: Yh ∓ $\sqrt{2F(1 - \alpha,\ 2,\ n - 2)}* S\{\widehat{Y}_h\}$

Yh = β0 + β1*Xh

$S\{\widehat{Y}_h\} = \sqrt{MSE} \times \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma(X_i - \bar{X})^2} \right] = \sqrt{0.01073} * (1/369 + (0.46 - 0.75)/0.90185$ = 0.03163

Plugging in the values:

Yh = β0 + β1*Xh*

= -0.1193 + 1.01886(0.46)= 0.34875*

Calculating the confidence band:

Upper = 0.34875 +0.07748 = 0.42623

Lower = 0.34875 - 0.07748 = 0.27127

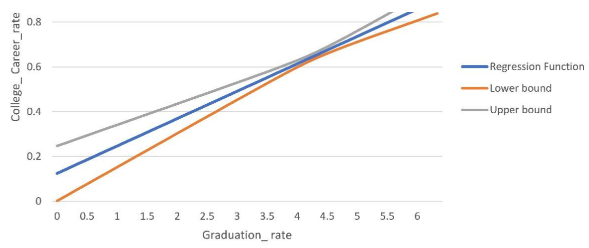
Therefore, the 95% confidence band for a Xh value of 0.46 lies between the values (0.27127, 0.4263).

Table 2: Limits for Working-Hotelling confidence bands.

graduation_rate(Xh)	Y(Xh)	S{Yh}	Lower bound	Upper bound
0.24	0.124596	0.050055	0.00198614	0.247206657
0.26	0.144974	0.048125	0.02709219	0.262855008
0.28	0.165351	0.046195	0.05219583	0.278505768
0.3	0.185728	0.044267	0.07729675	0.29415925
0.32	0.206105	0.04234	0.10239457	0.309815827
0.34	0.226482	0.040414	0.12748886	0.325475942
0.36	0.24686	0.03849	0.15257908	0.341140125
0.38	0.267237	0.036568	0.17766458	0.356809019
0.4	0.287614	0.034648	0.20274459	0.372483408
0.42	0.307991	0.032731	0.22781814	0.388164258
0.44	0.328368	0.030816	0.25288402	0.403852777
0.46	0.348746	0.028906	0.27794071	0.419550486
0.48	0.369123	0.027	0.30298626	0.435259338
0.5	0.3895	0.0251	0.32801813	0.45098187
0.52	0.409877	0.023207	0.35303296	0.466721443
0.54	0.430254	0.021322	0.3780262	0.482482597
0.56	0.450632	0.019449	0.4029916	0.498271604
0.58	0.471009	0.017591	0.42792023	0.514097365
0.6	0.491386	0.015753	0.45279911	0.529972887
0.62	0.511763	0.013944	0.47760855	0.545917851
0.64	0.53214	0.012175	0.50231757	0.561963235
0.66	0.552518	0.010468	0.52687522	0.57815998
0.68	0.572895	0.008859	0.55119386	0.594595741
0.7	0.593272	0.007412	0.57511714	0.61142686
0.72	0.613649	0.006239	0.59836732	0.628931084
0.74	0.634026	0.005519	0.62050819	0.647544608

0.76	0.654404	0.005435	0.64109132	0.667715883
0.78	0.674781	0.006013	0.66005121	0.689510388
0.8	0.695158	0.007094	0.67778061	0.712535386
0.82	0.715535	0.008488	0.69474445	0.736325951
0.84	0.735912	0.010065	0.71125868	0.760566122
0.86	0.75629	0.011752	0.72750374	0.785075458
0.88	0.776667	0.013507	0.74358033	0.809753266
0.9	0.797044	0.015308	0.75954638	0.834541623
0.92	0.817421	0.01714	0.7754367	0.8594057
0.94	0.837798	0.018994	0.79127321	0.884323593
0.96	0.858176	0.020864	0.80707024	0.909280961
0.98	0.878553	0.022746	0.82283753	0.934268069
1	0.89893	0.024637	0.8385819	0.959278103

Figure 6: Regression plot of (predictor) vs (response) with 95% confidence bands.



From Figure 6, we can observe that the regression line lies between the 95% confidence bands. This helps us conclude that for different values of Xh, we are 95% confident that the Yh value lies between its respective lower and upper bounds.

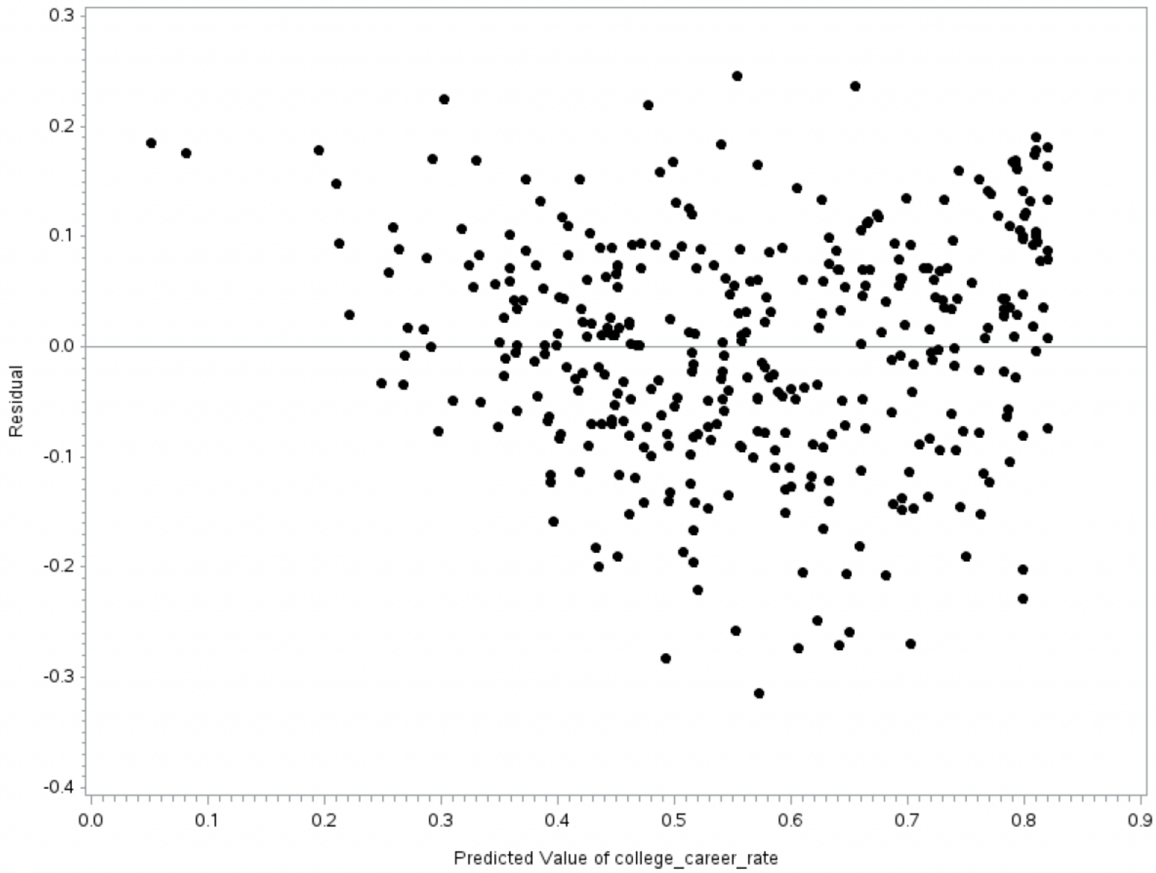
Now, we'll calculate s{predmean} by dividing Prediction Interval values by 39 (we have 39 observations of new Xh).

$(\frac{0.41960426}{39}, \frac{0.82627578}{39})$ = (0.01075908 , 0.02118656)

4. Model Assumptions

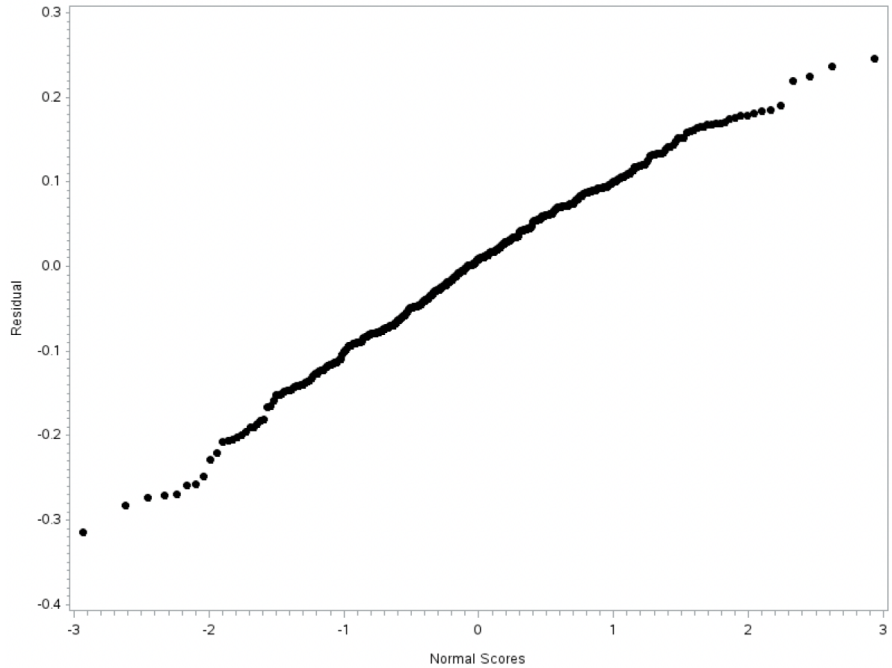
Before we move further with our regression analysis, we need to check if the linear model is appropriate, the residuals are uncorrelated, check for normality, no outliers, and for constant variance.

Figure 7: Residual Vs. Predicted value of college_career_rate



In Figure 7, we see a random point cloud and there is no curvature, so our linear model is appropriate. Furthermore, there is no funnel shape so constant variance is OK. Additionally, we don't see any y-outlier

Figure 8: Residual Vs. Normal Scores



In Figure 8, It is mostly linear so it means that the normality is satisfied.

Normality test with alpha (**α**) = 0.10 :

Hypothesis Tests:

H_0: Normality is OK

VS.

H_1: Normality is violated

Table 3: Pearson Correlation Coefficients

Pearson Correlation Coefficients, N = 369 Prob > |r| under H0: Rho=0		
	e	enrm
e Residual	1.00000	0.99560 <.0001
enrm Normal Scores	0.99560 <.0001	1.00000

Table 3 shows a normal score of 0.99560 so our $\hat{\rho}$= 0.99560.

Testing conditions:

If $\widehat{\rho} < c(\alpha, n)$, we reject H_0

Else we fail to reject H_0

From Critical values for the coefficient of correlation table, c(0.10, 369) = 0.989

Here $\widehat{\rho}$= 0.99560 > c(0.10, 369) = 0.989, so we fail to reject H_0.

Conclusion: We are 90% confident that Normality is Ok, so the normality plot agrees with normality test. Therefore, normality is satisfied.

Constant variance test.

Because the normality is satisfied we're conducting the Breush-Pagan test with alpha (**α**) = 0.05 :

Hypothesis Test:
$H_0: \gamma_1 = 0$
VS.
$H_1: \gamma_1 \neq 0$

Table 4: ANOVA table for squared residuals as the dependent variable

The REG Procedure
Model: MODEL1
Dependent Variable: e_sq Squared Residual

Number of Observations Read	440
Number of Observations Used	369
Number of Observations with Missing Values	71

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00050854	0.00050854	2.37	0.1249
Error	367	0.07891	0.00021502		
Corrected Total	368	0.07942			

Root MSE	0.01466	R-Square	0.0064
Dependent Mean	0.01068	Adj R-Sq	0.0037
Coeff Var	137.35365		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.00501	0.00376	1.33	0.1834
graduation_rate		1	0.00753	0.00490	1.54	0.1249

From Table 4, we can see the SSR*= 0.00050854.

Testing conditions:

If $\chi^2_{BP} > \chi^2(1-\alpha;1)$, we reject H_0

Else, we fail to reject H_0

$$\chi^2_{BP} = \frac{(SSR^*/2)}{(SSE/n)^2}$$
$$= \frac{(0.00050854/2)}{(3.93938/369)^2}$$
$$= \frac{0.000025427}{0.000113973}$$
$$= 0.2231$$

$\chi^2(1-\alpha;1) = \chi^2(0.95;1) = 3.84$

Since $\chi^2_{BP} = 0.2231 < \chi^2(1-\alpha;1) = 3.84$, we fail to reject H_0.

Conclusion:

We are 95% confident that the variance is constant, so the residual vs college_career_rate plot agrees with Breush-Pagan test. Therefore, we have constant variance.

This dataset is not time-related, so we don't need to use the time-series plots for our analysis.

5. Final discussion:

In conclusion, our simple linear model is $\widehat{College\_career\_rate} = -0.19928 + 1.0188$ graduation rate. The linear association between the dependent variable (college_career_rate) and the predictor variable (graduation_rate) is prominent, and the regression is significant.

Based on the model assumptions and hypothesis testing, we concluded that our linear model is appropriate. the normality is verified using the normality test and the assumption that variance is constant was proved accurate by using the Breusch Pagan test.

For future analysis, we can use other predictor variables such as attendence_rate and pct_stu_enough_variety. We can plot each predictor variable with our dependent variable (college_career_rate), explore the interaction terms, and further test our hypothesis before we add the new predictor variable to the model. The SLR model, that has been built can be used to predict future values of College_career_rate using graduation_rate exclusively.

III. Multiple Linear Regression

We used the same dataset for Multiple Linear Regression, but since there were a lot of missing data points, we performed imputation. Multiple Linear Regression is based on the imputed dataset.

1. Multiple Linear Regression Model

Multiple linear regression takes the form: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$, where, Y_i is the response variable, which is the college_career_rate, and X_{i1}, X_{i2}, X_{i3}, X_{i4} are the predictor variables: graduation_rate, attendance_rate, pct_stu_safe, pct_stu_enough_variety respectively.

β_0, β_1, β_2, β_3, β_4 are the regression coefficients, where they represent the intercept and slope of the regression line respectively. ε_i is the random error term associated with the model.

Table 5: ANOVA TABLE

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Number of Observations Read	440
Number of Observations Used	440

X'X Inverse, Parameter Estimates, and SSE

Variable	Intercept	graduation_rate	attendance_rate	pct_stu_safe	pct_stu_enough_variety	college_career_rate
Intercept	0.7027087045	0.1978434708	-0.63344042	-0.290320317	-0.069271027	-0.667515315
graduation_rate	0.1978434708	0.2337924912	-0.280866993	-0.158014007	0.0053643753	0.7458806116
attendance_rate	-0.63344042	-0.280866993	0.9298214205	0.0409233665	0.0016698845	0.5808285903
pct_stu_safe	-0.290320317	-0.158014007	0.0409233665	0.554338222	-0.124022869	0.2216787501
pct_stu_enough_variety	-0.069271027	0.0053643753	0.0016698845	-0.124022869	0.2206338853	-0.025661773
college_career_rate	-0.667515315	0.7458806116	0.5808285903	0.2216787501	-0.025661773	3.9018346395

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9.34145	2.33536	260.36	<.0001
Error	435	3.90183	0.00897		
Corrected Total	439	13.24328			

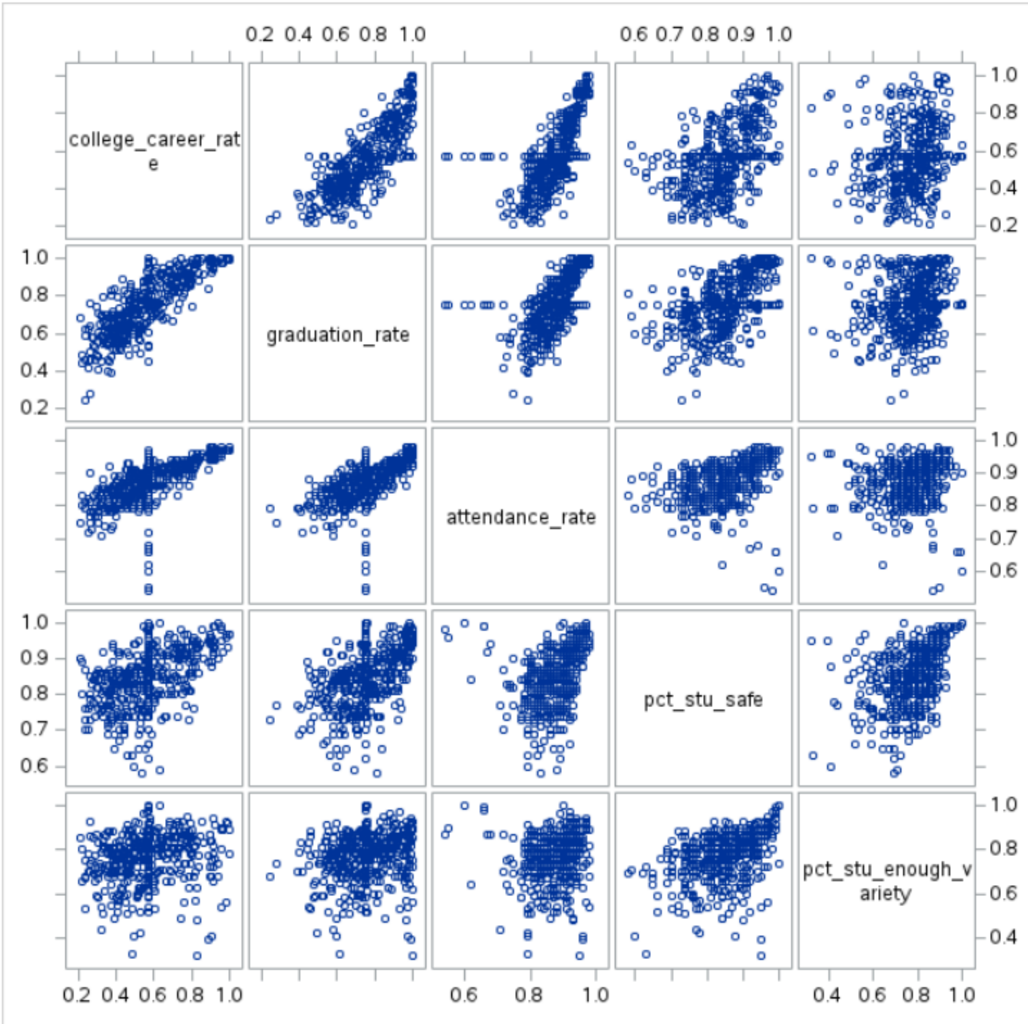
Root MSE	0.09471	R-Square	0.7054
Dependent Mean	0.56672	Adj R-Sq	0.7027
Coeff Var	16.71178		

The point estimator for b_0 is −0.66752, b_1 is 0.74588, b_2 is 0.58083, b_3 is 0.22168, b_4 is -0.02566, b_5 is 0.3902. So our full model is as follows:

$\widehat{College\_career\_rate}$ = **−0.66752 + 0.74588*graduation_rate + 0.58083*attendence_rate + 0.22168*pct_stu_safe - 0.02566*pct_stu_enough_variety**

The model obtained an R-Square value of 0.7054, which indicates a linear association between college_career_rate and graduation_rate.

Figure 9: Matrix scatter plot



- We can observe that there is a strong positive linear association between graduation_rate & college_career_rate with few potential outliers.
- There appears to be a moderately positive linear association between attendence_rate & college_career_rate.
- A weak positive linear connection is evident between pct_stu_enough_variety and college_career_rate, similar to the relationship between pct_stu_safe and college_career_rate.
- A strong positive correlation is observed between graduation_rate & attendance_rate while a weak positive linear association among the remaining predictor vs predictor.
- From the matrix scatter plot, with respect to college_career_rate, we can observe that graduation_rate and attendence_rate have low variance, while predictors pct_stu_safe and pct_stu_enough_variety have high variance.

Table 6: Pairwise correlation among variables

Pearson Correlation Coefficients, N = 440					
	college_career_rate	graduation_rate	attendance_rate	pct_stu_safe	pct_stu_enough_variety
college_career_rate	1.00000	0.82012	0.65382	0.49629	0.16148
graduation_rate	0.82012	1.00000	0.64787	0.53114	0.18388
attendance_rate	0.65382	0.64787	1.00000	0.30386	0.09996
pct_stu_safe	0.49629	0.53114	0.30386	1.00000	0.39368
pct_stu_enough_variety	0.16148	0.18388	0.09996	0.39368	1.00000

- We can see that among the 4 predictors, graduation_rate has a high correlation with college_career_rate while pct_stu_enough_variety has the least correlation.
- Both attendence_rate and pct_stu_safe are moderately correlated to the response variable college_career_rate.
- Among the predictors variables, graduation_rate is moderately correlated with attendence_rate. On the contrary, pct_stu_safe is less correlated with pct_stu_enough_variety, graduation_rate, and attendance_rate.
- Pct_stu_enough_variety is the least correlated with any other predictor variables.

Discuss potential complications:

1. Looking at the correlation values, we can see that the correlation between the predictor variables vs. predictor variables are always less than 0.65. This suggests a weak or moderate correlation among predictor variables. Thus less chance of Multicollinearity among the predictor variables.

2. Schools with low college_career_rate show some outlier behavior. The graduation rate has few outliers. Moderate numbers of outliers in attendance_rate and high number of outliers in pct_stu_safe & pct_stu_enough_variety can affect the model accuracy.

2. Preliminary Multiple Linear regression model analysis

Table 7: Variance Inflation Table

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

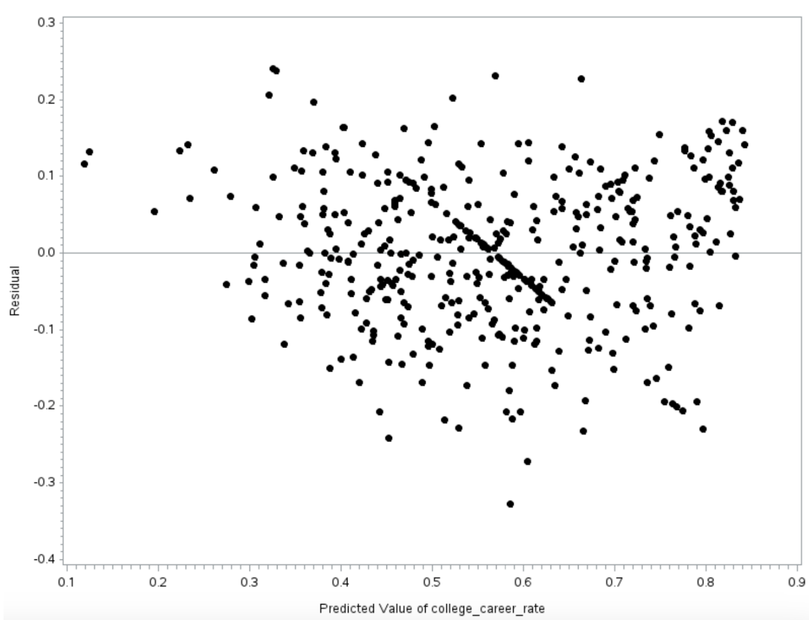
Number of Observations Read	440
Number of Observations Used	440

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	-0.66752	0.07939	-8.41	<.0001	141.31457	0
graduation_rate	1	0.74588	0.04579	16.29	<.0001	8.90746	2.18868
attendance_rate	1	0.58083	0.09132	6.36	<.0001	0.34237	1.73012
pct_stu_safe	1	0.22168	0.07051	3.14	0.0018	0.08863	1.59958
pct_stu_enough_variety	1	-0.02566	0.04449	-0.58	0.5643	0.00298	1.18466

The variance inflation for graduation_rate, attendance_rate, pct_stu_safe, and pct_stu_enough_variety are all less than 5. Max $VIF_{graduation\_rate}$ = 2.18868 is less than 5 and $(\overline{VIF})$ = 1.67576, is not much bigger than 1. So, we conclude that serious multicollinearity is not a problem.

Figure 10: Residual vs Predicted value of college_career_rate



In Figure 10, we see a random point cloud and there is no curvature, so our linear model is appropriate. Furthermore, there is no funnel shape so constant variance is OK. Additionally, we don't see any y-outlier in this figure.

Table 8: Check outliers, Leverage, and Influence

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Output Statistics										
						DFBETAS				
Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	Intercept	graduation_rate	attendance_rate	pct_stu_safe	pct_stu_enough_variety
1	-0.0454	-0.4594	0.1448	1.2774	-0.1890	0.1574	0.0964	-0.1360	-0.0982	-0.0153
2	-0.001738	-0.0215	0.4293	1.9607	-0.0186	-0.0098	-0.0064	0.0172	-0.0019	0.0007
3	0.0964	1.0998	0.3115	1.4192	0.7398	0.5476	-0.0890	-0.2298	-0.0791	-0.4793
4	-0.1453	-1.6157	0.2524	1.1219	-0.9388	-0.4283	-0.5736	0.1205	0.8807	-0.4572
5	-0.0464	-0.4685	0.1394	1.2682	-0.1885	-0.0036	0.0807	-0.0084	-0.1262	0.1719
6	-0.0349	-0.3323	0.0381	1.1488	-0.0662	0.0248	0.0008	-0.0239	0.0106	-0.0385
7	-0.1318	-1.2905	0.0551	0.9834	-0.3117	0.0533	0.1909	-0.1921	-0.0494	0.0905
8	-0.2809	-3.0134	0.0853	0.4787	-0.9200	0.2525	-0.1650	-0.2890	0.4413	-0.7716
9	0.0789	0.7571	0.0408	1.0934	0.1561	0.0077	0.0346	0.0427	-0.0963	0.0757
10	0.0299	0.2924	0.0872	1.2140	0.0904	0.0182	-0.0417	0.0254	-0.0326	0.0299
11	0.0129	0.1225	0.0261	1.1470	0.0201	-0.0033	0.0004	0.0050	-0.0044	0.0087
12	0.0589	0.6362	0.2454	1.4164	0.3628	-0.0240	0.0034	0.0173	0.1730	-0.3002
13	0.0199	0.1972	0.1084	1.2495	0.0688	-0.0014	0.0233	-0.0076	0.0193	-0.0409
14	0.1356	1.3614	0.0984	1.0097	0.4499	-0.3014	0.1057	0.1139	0.1957	-0.0229
15	0.0125	0.1184	0.0245	1.1453	0.0188	0.0080	-0.0003	-0.0068	-0.0000	-0.0025
16	0.0623	0.6318	0.1436	1.2488	0.2587	0.1244	0.0436	0.0243	-0.1558	-0.0316
17	0.0382	0.3749	0.0900	1.2101	0.1179	-0.0824	0.0190	0.0518	0.0022	0.0670
18	-0.0105	-0.1201	0.3370	1.6850	-0.0856	-0.0296	-0.0258	0.0705	-0.0128	-0.0097
19	0.0735	0.7205	0.0800	1.1469	0.2125	0.0571	-0.1604	0.0319	0.0392	-0.0706
20	0.0235	0.2267	0.0573	1.1800	0.0559	-0.0310	-0.0213	0.0159	0.0288	0.0083
21	0.1495	1.4756	0.0595	0.9344	0.3712	-0.1277	0.1142	0.0736	0.0869	-0.1131
22	-0.2661	-2.8613	0.1050	0.5326	-0.9802	0.0888	-0.4739	0.1389	-0.2174	0.4207
23	0.2533	2.5719	0.0260	0.5704	0.4206	-0.1087	-0.0940	0.1771	-0.0130	0.0948
24	0.0365	0.3574	0.0869	1.2078	0.1103	0.0334	-0.0336	-0.0020	-0.0410	0.0537
25	0.006667	0.0630	0.0240	1.1459	0.0099	-0.0023	-0.0014	0.0037	-0.0004	0.0019
26	-0.0659	-0.6340	0.0498	1.1252	-0.1451	-0.0680	0.0702	0.0256	-0.0327	0.0772
27	-0.0446	-0.4255	0.0364	1.1376	-0.0827	-0.0493	0.0105	0.0417	-0.0036	0.0199
28	-0.2114	-2.1721	0.0851	0.7349	-0.6624	0.1355	-0.2963	-0.0203	-0.1169	0.2349
29	0.0215	0.2062	0.0516	1.1741	0.0481	0.0188	0.0010	-0.0260	0.0188	-0.0282
30	-0.0529	-0.5151	0.0728	1.1711	-0.1444	-0.0488	0.1018	-0.0100	-0.0194	0.0420
31	0.0272	0.2699	0.1150	1.2539	0.0973	0.0535	0.0465	-0.0806	-0.0233	0.0281
32	-0.1260	-1.2367	0.0627	1.0063	-0.3198	-0.1825	0.1069	0.0352	0.1125	-0.0289
33	0.0395	0.3802	0.0550	1.1648	0.0917	-0.0362	0.0510	0.0062	-0.0060	0.0317
34	-0.0269	-0.2606	0.0663	1.1891	-0.0694	-0.0544	-0.0005	0.0232	0.0265	0.0183
35	-0.0646	-0.6191	0.0425	1.1190	-0.1305	0.0243	0.0184	-0.0377	0.0329	-0.0802
36	0.1235	1.2618	0.1334	1.0809	0.4950	0.1490	0.3909	-0.0813	-0.3691	0.1652
37	0.0719	0.7154	0.1088	1.1849	0.2499	0.0741	0.1065	0.0051	-0.2211	0.1676
38	0.1175	1.1367	0.0410	1.0096	0.2351	0.0588	0.0807	-0.0645	-0.1155	0.1481
39	-0.0846	-0.9013	0.2176	1.3052	-0.4754	0.0405	0.4310	-0.0925	-0.2949	0.1815
40	-0.0765	-0.7498	0.0804	1.1419	-0.2217	0.0675	-0.1445	0.0040	0.0613	-0.1322
41	0.0712	0.7117	0.1168	1.1964	0.2588	-0.0359	-0.1834	0.0148	0.1359	-0.0110
42	0.1187	1.1723	0.0778	1.0404	0.3406	-0.2188	0.0885	0.1004	0.1195	-0.0120
43	-0.0943	-0.9048	0.0336	1.0559	-0.1687	0.0747	0.0559	-0.0643	-0.0905	0.0533
44	-0.0785	-0.7677	0.0756	1.1325	-0.2195	0.1164	-0.0902	-0.0373	-0.0541	0.0018
45	0.0392	0.3771	0.0560	1.1663	0.0918	-0.0583	-0.0467	0.0477	0.0470	0.0047
46	0.0118	0.1119	0.0291	1.1509	0.0194	0.0061	0.0040	0.0007	-0.0088	0.0012
47	0.004897	0.0478	0.0847	1.2221	0.0145	0.0003	-0.0089	0.0044	-0.0013	0.0043
48	0.0959	0.9496	0.0924	1.1140	0.3031	-0.2202	0.0160	0.1114	0.0672	0.1417
49	0.0510	0.4930	0.0618	1.1602	0.1265	0.0358	0.0405	0.0143	-0.1003	0.0623
50	0.006880	0.0662	0.0576	1.1867	0.0164	-0.0057	-0.0102	0.0090	0.0086	-0.0083

Check for x-outlier

$h_{ii} = \frac{2p}{n} = \frac{2*5}{50} = 0.2$

Observations (2,3,4,12,18,39) have Leverage values > h_{ii} (0.2), so we have 6 x-outliers.

Check for y-outlier

t(1-$\frac{\alpha}{2n}$; n-p-1) = t(1-$\frac{0.05}{2*50}$; 50-5-1)

From t-distribution table t(0.9995;44) = 3.5328

Since $|t|$ < t(0.9995;44) = 3.5328, we don't have y-outliers.

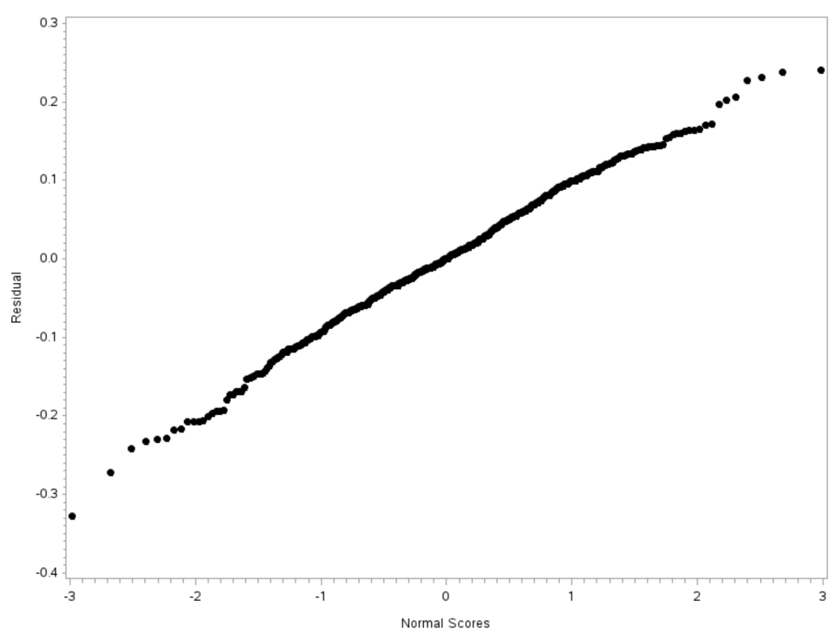
Check for the influence

DFFITS = $2\sqrt{\frac{p}{n}}$

$= 2\sqrt{\frac{5}{50}}$ = 0.63246

At observation (3) x has influence on $\widehat{y}$, but there is no influence between y and $\widehat{y}$.

Figure 11: Residual vs Normal Scores



In figure 11, we see that the plot is mostly linear so it means that the normality is satisfied.

Normality test with alpha (α) = 0.10 :

Hypothesis Tests:

H_0: Normality is OK

VS.

H_1: Normality is violated

Table 9: Pearson Correlation Coefficients

Pearson Correlation Coefficients, N = 440 Prob > |r| under H0: Rho=0		
	e	enrm
e Residual	1.00000	0.99738 <.0001
enrm Normal Scores	0.99738 <.0001	1.00000

Testing conditions:

If $\hat{\rho} < c(\alpha, n)$, we reject H_0

Else we fail to reject H_0

From Critical values for coefficient of correlation table, c(0.10, 440) = 0.989

Here $\hat{\rho} = 0.99738 > c(0.10, 440) = 0.989$, so we fail to reject H_0.

Conclusion:

We are 90% confident that Normality is Ok. So, our test agrees with what we saw in the normality plot (figure 11)

Since normality is OK, we will use the Breusch-Pagan test to conduct the constant variance.

Table 10: ANOVA table for squared residuals as the dependent variable

The REG Procedure
Model: MODEL1
Dependent Variable: e_sq Squared Residual

Number of Observations Read	440
Number of Observations Used	440

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.00132	0.00033056	2.07	0.0844
Error	435	0.06961	0.00016003		
Corrected Total	439	0.07094			

From table 10, we can see the SSR*= 0.00132.

Testing conditions:

If $\chi^2_{BP} > \chi^2(1-\alpha;4)$, we reject H_0

Else, we fail to reject H_0

$$\chi^2_{BP} = \frac{(SSR^*/2)}{(SSE/n)^2}$$
$$= \frac{(0.00132/2)}{(3.90183/440)^2}$$
$$= \frac{0.00066}{0.00008}$$
$$= 8.25$$

$\chi^2(1-\alpha;4) = \chi^2(0.95;4) = 9.49$

Since $\chi^2_{BP} = 8.25 < \chi^2(1-\alpha;4) = 9.49$, we fail to reject H_0.

Conclusion:

We are 95% confident that the variance is constant. So, the residual vs college_career_rate plot (figure 10) agrees with our Breush-Pagan test. Therefore, we have constant variance.

3. **Explore Interaction terms:**

Table 11: Correlations between Standardized Predictors vs Interaction terms and Standardized Interaction terms

Correlations between Interaction terms and Predictors

	college_career_rate	graduation_rate	attendance_rate	pct_stu_safe	pct_stu_enough_variety
x1x2	0.84229	0.97187	0.80447	0.51109	0.16769
x1x3	0.84229	0.97187	0.80447	0.51109	0.16769
x1x4	0.84229	0.97187	0.80447	0.51109	0.16769
x2x3	0.69790	0.71371	0.76553	0.84356	0.31226
x2x4	0.42011	0.43497	0.53800	0.45571	0.88887
x3x4	0.34652	0.37885	0.20191	0.74616	0.90104

Correlations between Standardized Interaction terms and Predictors

	college_career_rate	graduation_rate	attendance_rate	pct_stu_safe	pct_stu_enough_variety
stdx1x2	0.15272	-0.00966	0.00773	0.12654	-0.10147
stdx1x3	0.20464	0.05317	0.12008	0.02369	-0.05631
stdx1x4	-0.08173	-0.10997	-0.08395	-0.04909	0.07321
stdx2x3	0.18137	0.09953	0.42976	-0.02145	-0.16694
stdx2x4	-0.05372	-0.07344	0.17308	-0.15361	-0.05888
stdx3x4	-0.04265	-0.04449	-0.15914	-0.09584	0.01265

We can observe the correlation between the interaction terms x_1x_2, x_1x_3, and x_1x_4, and predictors and can observe most of the correlation values are greater than 0.8 which shows that they are strongly correlated to each other, x_2x_3 have a correlation value of 0.7. x_2x_4 and x_3x_4 are moderately correlated to the predictors.

We can observe that the correlation values have decreased after standardization of the interaction terms with 0.42 being the maximum value, which tells that the correlation among the interaction terms and the predictors is reduced.

Figure 12: Residual vs graduation_rate*attendence_rate

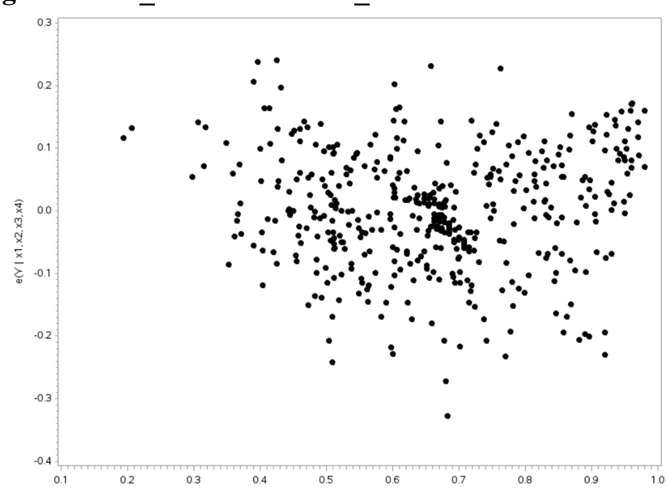


Figure 12 shows the residual vs graduation_rate*attendence_rate and we can observe that most of the data points are scattered with a slightly linear relationship between them so we may or may not add graduation_rate*attendence_rate as the predictor to the model.

Figure 13: Residual vs Standardized graduation_rate*attendence_rate

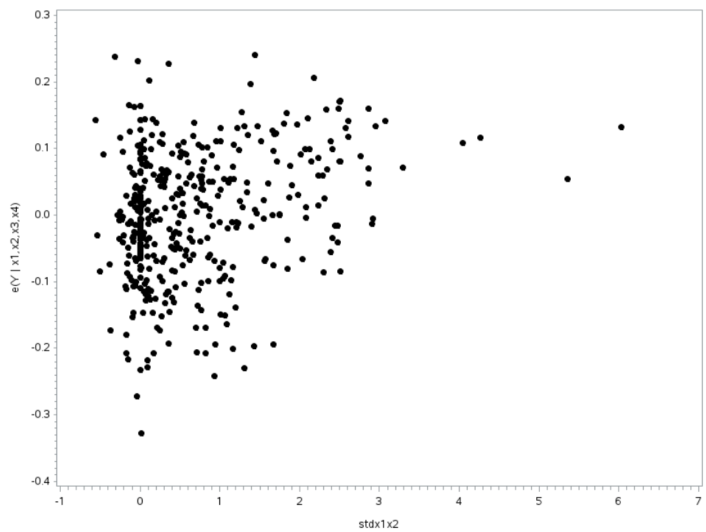


Figure 13 shows residual vs Standardized graduation_rate*attendence_rate, Here we can observe that there are fewer scatter data points compared to the original interaction and has a positive slope so we can add Standardized graduation_rate*attendence_rate as the predictor to the model.

Figure 14 $residual\ _{Full}$**vs** $residual\ _{Reduced}$**(graduation_rate*attendence_rate)**

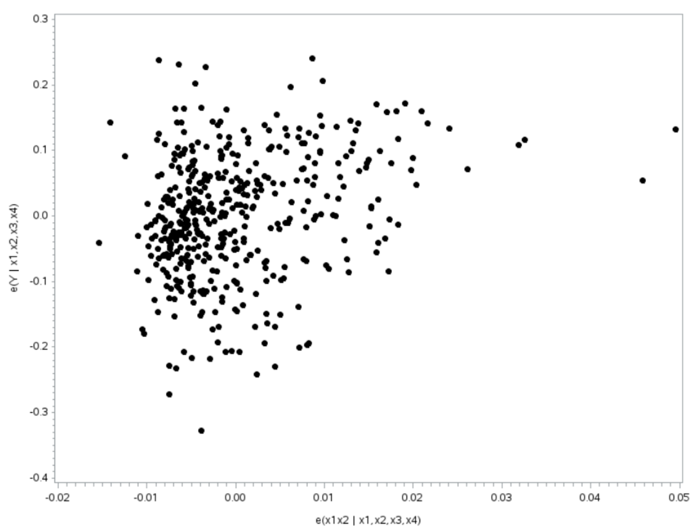


Figure 14 shows $residual\ _{Full}$vs $residual\ _{Reduced}$(graduation_rate*attendence_rate), we can observe that the points are more scattered and do not have a linear trend compared to the previous model of standardized values so we cannot consider this as a predictor variable.

Figure 15: Residual vs. attendance_rate*pct_stu_enough_variety

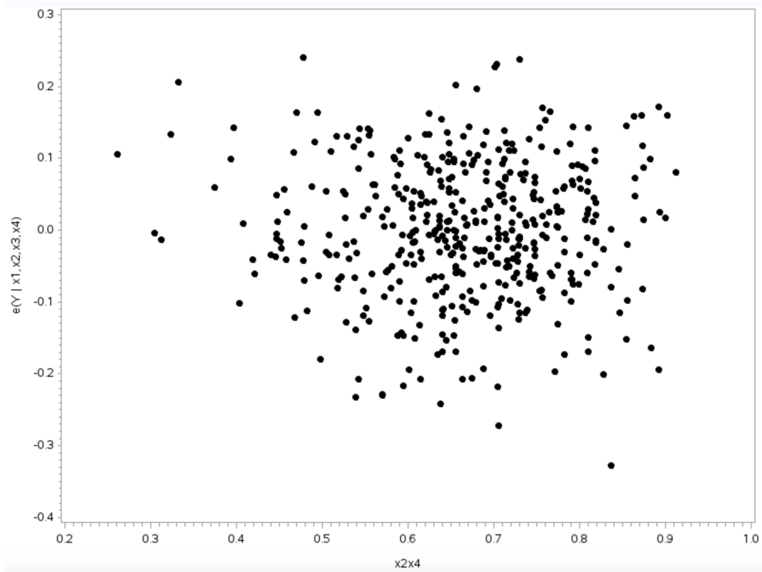


Figure 15 shows the residual vs. attendance_rate*pct_stu_enough_variety (interaction term) plot and there is a random point cloud. So, it is not helpful to add attendance_rate*pct_stu_enough_variety as a predictor variable in our model.

Figure 16: Residual vs. Standardized attendance_rate*pct_stu_enough_variety

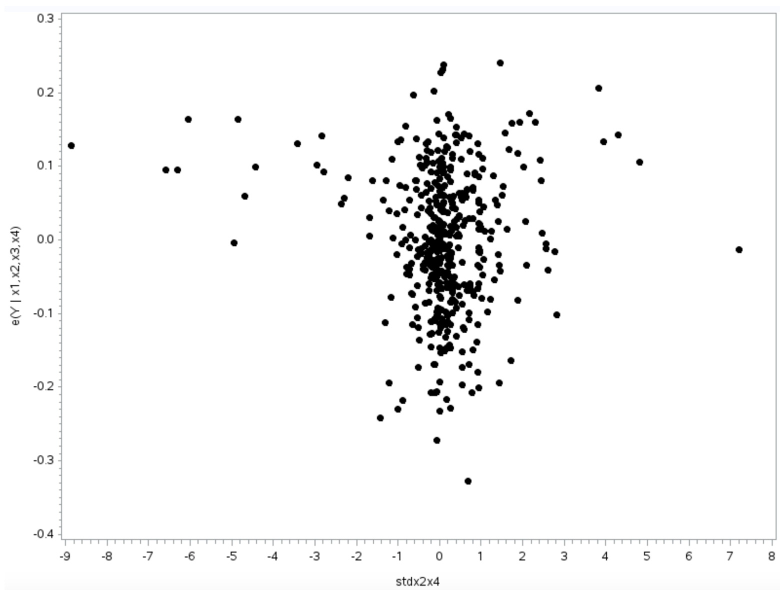


Figure 16 shows the residual vs. standardized attendance_rate*pct_stu_enough_variety (standardized interaction term) plot and there is a line in the middle and a random point cloud around it. It looks slightly linear. It may or may not be linear so we can add standardized attendance_rate*pct_stu_enough_variety to our model.

Figure 17:$residual_{Full}$ **vs** $residual_{Reduced}$**(attendance_rate*pct_stu_enough_variety)**

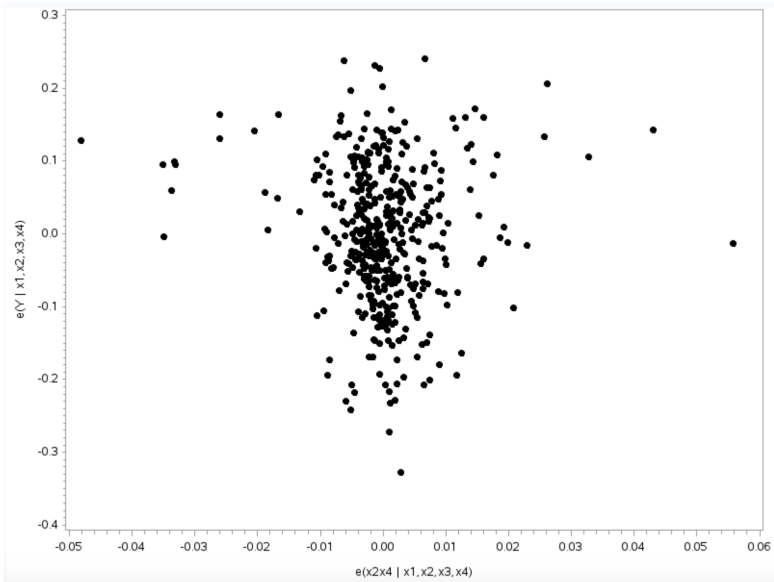


Figure 17 shows residual(college_career_rate | all predictor variables) vs. residual(attendance_rate*pct_stu_enough_variety | all predictor variables) and it does not look linear. We should not add attendance_rate*pct_stu_enough_variety to our model.

4. Best Subset Regression Method:

Table 12: Adjusted R-Square Selection Method Output

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Adjusted R-Square Selection Method

Number of Observations Read	440
Number of Observations Used	440

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
1	0.6719	0.6726	107.6658	-2028.7394	-2020.56584	graduation_rate
1	0.4262	0.4275	514.7145	-1782.8326	-1774.65906	attendance_rate

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
2	0.6972	0.6986	66.4820	-2063.1552	-2050.89488	graduation_rate stdx1x3
2	0.6971	0.6985	66.7359	-2062.9321	-2050.67174	graduation_rate attendance_rate

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
3	0.7214	0.7233	27.4250	-2098.8174	-2082.47027	graduation_rate attendance_rate stdx1x2
3	0.7173	0.7192	34.2650	-2092.3149	-2075.96780	graduation_rate attendance_rate stdx1x3

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
4	0.7269	0.7294	19.3758	-2106.5485	-2086.11463	graduation_rate attendance_rate stdx1x2 stdx2x3
4	0.7263	0.7288	20.3971	-2105.5496	-2085.11577	graduation_rate attendance_rate stdx1x2 stdx2x4

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
5	0.7355	0.7385	6.2445	-2119.6193	-2095.09861	graduation_rate attendance_rate stdx1x2 stdx1x3 stdx2x3
5	0.7295	0.7326	16.0179	-2109.8262	-2085.30552	graduation_rate attendance_rate stdx1x2 stdx1x3 stdx2x4

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
6	0.7373	0.7409	4.1879	-2121.7490	-2093.14158	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx2x3
6	0.7351	0.7387	7.8454	-2118.0238	-2089.41642	graduation_rate attendance_rate stdx1x2 stdx1x3 stdx2x3 stdx2x4

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
7	0.7370	0.7412	5.7670	-2120.1797	-2087.48550	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx1x4 stdx2x3
7	0.7369	0.7411	5.9493	-2119.9931	-2087.29892	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx2x3 stdx2x4

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
8	0.7368	0.7416	7.0594	-2118.9047	-2082.12372	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx3x4
8	0.7365	0.7413	7.5933	-2118.3576	-2081.57663	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx2x3 stdx2x4 stdx3x4

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
9	0.7362	0.7417	9.0036	-2116.9619	-2076.09420	graduation_rate attendance_rate pct_stu_safe pct_stu_enough_variety stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx3x4
9	0.7362	0.7416	9.0577	-2116.9065	-2076.03877	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx2x4 stdx3x4

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
10	0.7351	0.7411	11.0000	-2109.0460	-2064.11652	graduation_rate attendence_rate pct_stu_safe pct_stu_enogh_variety stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx2x4 stdx3x4

The above tables are the outputs for the best subset selection method. We choose the potentially best models based on the R-Squared, C(p), AIC and SBC values. Out of the 10 models that we have, we can say that model 6 and 7 are potentially best because they have a higher R-Squared value compared to the other models. These models also have low C(p), AIC and SBC values. We eliminate the models 9 and 10 because their R-squared values have decreased from 0.7362 to 0.7351. Furthermore, their C(p), AIC and SBC values have increased.

Table 13: Summary table of Adjusted R-Square Method

Number in Model	Adjusted R-Square	R-Square	C(p)	AIC	SBC	Variables in Model
6	0.7373	0.7409	4.1879	-2121.7490	-2093.14158	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx2x3
7	0.7370	0.7412	5.7670	-2120.1797	-2087.48550	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx1x4 stdx2x3
7	0.7369	0.7411	5.9493	-2119.9931	-2087.29892	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx2x3 stdx2x4
8	0.7368	0.7416	7.0594	-2118.9047	-2082.12372	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx3x4
7	0.7368	0.7410	6.0679	-2119.8718	-2087.17758	graduation_rate attendance_rate pct_stu_safe pct_stu_enough_variety stdx1x2 stdx1x3 stdx2x3
7	0.7368	0.7410	6.0804	-2119.8589	-2087.16474	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx2x3 stdx3x4
8	0.7365	0.7413	7.5933	-2118.3576	-2081.57663	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx2x3 stdx2x4 stdx3x4
8	0.7364	0.7412	7.7015	-2118.2468	-2081.46581	graduation_rate attendance_rate pct_stu_safe pct_stu_enough_variety stdx1x2 stdx1x3 stdx1x4 stdx2x3
8	0.7364	0.7412	7.7662	-2118.1805	-2081.39949	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx2x4
8	0.7363	0.7411	7.8751	-2118.0690	-2081.28807	graduation_rate attendance_rate pct_stu_safe pct_stu_enough_variety stdx1x2 stdx1x3 stdx2x3 stdx2x4
8	0.7363	0.7411	7.9454	-2117.9971	-2081.21610	graduation_rate attendance_rate pct_stu_safe pct_stu_enough_variety stdx1x2 stdx1x3 stdx2x3 stdx3x4
9	0.7362	0.7417	9.0036	-2116.9619	-2076.09420	graduation_rate attendance_rate pct_stu_safe pct_stu_enough_variety stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx3x4
9	0.7362	0.7416	9.0577	-2116.9065	-2076.03877	graduation_rate attendance_rate pct_stu_safe stdx1x2 stdx1x3 stdx1x4 stdx2x3 stdx2x4 stdx3x4
9	0.7359	0.7413	9.5214	-2116.4312	-2075.56342	graduation_rate attendance_rate pct_stu_safe pct_stu_enough_variety stdx1x2 stdx1x3 stdx2x3 stdx2x4 stdx3x4

The above summary table shows that model 6 is the best model. It has the highest Adjusted R-Square value, lowest C(p), AIC and SBC values among the other models. Since we are picking two best models, model 7 is the next best model because it has a high R-Squared value and low C(p), AIC, and SBC values.

Stepwise Regression:

In stepwise selection, we start with none of the predictor variables and add one predictor variable with each step forward.

Table 14: Step-1 of Stepwise Selection Method

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Number of Observations Read	440
Number of Observations Used	440

Stepwise Selection: Step 1

Variable graduation_rate Entered: R-Square = 0.6726 and C(p) = 107.6658

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	8.90746	8.90746	899.82	<.0001
Error	438	4.33583	0.00990		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.16724	0.02492	0.44574	45.03	<.0001
graduation_rate	0.97544	0.03252	8.90746	899.82	<.0001

In STEP 1 of stepwise regression, we have graduation_rate as our first variable of the model, as shown in the above figure. P-value is less than the alpha value of 0.05, and this is the least alpha value among all predictors. We proceed to STEP 2 and add a new variable.

Table 15: Step 2 of Stepwise Selection Method

Stepwise Selection: Step 2

Variable stdx1x3 Entered: R-Square = 0.6986 and C(p) = 66.4820

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	9.25185	4.62593	506.47	<.0001
Error	437	3.99143	0.00913		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.17509	0.02397	0.48717	53.34	<.0001
graduation_rate	0.96523	0.03128	8.69725	952.22	<.0001
stdx1x3	0.02931	0.00477	0.34440	37.71	<.0001

In STEP 2, we have stdx1x3 as our next variable as shown in the above figure. P-value of stdx1x3 is less than the alpha value of 0.05, so we proceed to STEP 3 and add a new variable.

Table 16: Step 3 of Stepwise Selection Method

Stepwise Selection: Step 3

Variable attendance_rate Entered: R-Square = 0.7192 and C(p) = 34.2650

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9.52474	3.17491	372.26	<.0001
Error	436	3.71854	0.00853		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.50451	0.06268	0.55262	64.79	<.0001
graduation_rate	0.82012	0.03964	3.64979	427.94	<.0001
attendance_rate	0.50595	0.08945	0.27289	32.00	<.0001
stdx1x3	0.02635	0.00464	0.27491	32.23	<.0001

In STEP 3, we have attendence_rate as our next variable as shown in the above figure. Since p-value is less than alpha value we proceed to STEP 4 and add a new variable.

Table 17: Step 4 of Stepwise Selection Method

Stepwise Selection: Step 4

Variable stdx1x2 Entered: R-Square = 0.7264 and C(p) = 24.2999

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	9.62016	2.40504	288.75	<.0001
Error	435	3.62312	0.00833		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.52954	0.06238	0.60025	72.07	<.0001
graduation_rate	0.81982	0.03918	3.64712	437.88	<.0001
attendance_rate	0.52717	0.08861	0.29477	35.39	<.0001
stdx1x2	0.02123	0.00627	0.09542	11.46	0.0008
stdx1x3	0.01328	0.00600	0.04087	4.91	0.0273

In STEP 4, we have stdx_1x_2 as our next variable as shown in the above figure. Since the p-value is less than alpha value we proceed to STEP 5 and add a new variable.

Table 18: Step 5 of Stepwise Selection Method

Stepwise Selection: Step 5

Variable stdx2x3 Entered: R-Square = 0.7385 and C(p) = 6.2445

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	9.78011	1.95602	245.13	<.0001
Error	434	3.46317	0.00798		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.71128	0.07332	0.75100	94.11	<.0001
graduation_rate	0.76277	0.04041	2.84315	356.30	<.0001
attendance_rate	0.78109	0.10363	0.45330	56.81	<.0001
stdx1x2	0.02931	0.00640	0.16744	20.98	<.0001
stdx1x3	0.02498	0.00642	0.12067	15.12	0.0001
stdx2x3	-0.02526	0.00564	0.15994	20.04	<.0001

In STEP 5, we have stdx_2x_3 as our next variable as shown in the above figure. Since p-value is less than alpha value we proceed to STEP 6 and add a new variable.

Table 19: Step 6 of Stepwise Selection Method

Stepwise Selection: Step 6

Variable pct_stu_safe Entered: R-Square = 0.7409 and C(p) = 4.1879

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	9.81246	1.63541	206.40	<.0001
Error	433	3.43082	0.00792		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.78350	0.08133	0.73527	92.80	<.0001
graduation_rate	0.72538	0.04431	2.12300	267.94	<.0001
attendance_rate	0.77202	0.10336	0.44200	55.78	<.0001
pct_stu_safe	0.12942	0.06405	0.03235	4.08	0.0439
stdx1x2	0.02614	0.00657	0.12565	15.86	<.0001
stdx1x3	0.02576	0.00641	0.12779	16.13	<.0001
stdx2x3	-0.02350	0.00569	0.13521	17.06	<.0001

In STEP 6, we have pct_stu_safe as our next variable as shown in the above figure. Since all the variables have p-value less than alpha value we don't eliminate furthermore.

Table 20: summary of the forward selection method

Bounds on condition number: 2.5091, 76.623

All variables left in the model are significant at the 0.0500 level.

No other variable met the 0.0500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	graduation_rate		1	0.6726	0.6726	107.666	899.82	<.0001
2	stdx1x3		2	0.0260	0.6986	66.4820	37.71	<.0001
3	attendance_rate		3	0.0206	0.7192	34.2650	32.00	<.0001
4	stdx1x2		4	0.0072	0.7264	24.2999	11.46	0.0008
5	stdx2x3		5	0.0121	0.7385	6.2445	20.04	<.0001
6	pct_stu_safe		6	0.0024	0.7409	4.1879	4.08	0.0439

Table 20 displays the summary of the forward selection method. This provides us with the information that an effective predictive model is built using the six variables that were selected in the stepwise regression process. Only the variables with P-value less than alpha value of 0.05 are used for the reduced model.

Backward Deletion:

In Backward deletion, we start with all of the predictor variables and remove one variable at a time based on the p-value. The variable with the largest p-value is removed first and so on. This method includes step 0 with all the variables present in the model.

Table 21: Step 0 of Backward Elimination Method

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Number of Observations Read	440
Number of Observations Used	440

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7417 and C(p) = 11.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	9.82193	0.98219	123.16	<.0001
Error	429	3.42135	0.00798		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.79135	0.08504	0.69054	86.59	<.0001
graduation_rate	0.71722	0.04523	2.00517	251.43	<.0001
attendance_rate	0.78390	0.10656	0.43158	54.12	<.0001
pct_stu_safe	0.14221	0.07012	0.03281	4.11	0.0432
pct_stu_enough_variety	-0.01049	0.04369	0.00045985	0.06	0.8103
stdx1x2	0.02677	0.00674	0.12585	15.78	<.0001
stdx1x3	0.02552	0.00743	0.09409	11.80	0.0007
stdx1x4	-0.00513	0.00711	0.00416	0.52	0.4706
stdx2x3	-0.02353	0.00835	0.06333	7.94	0.0051
stdx2x4	0.00044049	0.00733	0.00002882	0.00	0.9521
stdx3x4	0.00380	0.00454	0.00558	0.70	0.4035

In STEP 0 of backward deletion, we consider all the predictors as shown in the image above. Among all the predictors, we have a high p-value for stdx_2x_4. Hence, in the next step we eliminate it.

Table 22:Step 1 of Backward Elimination Method

Backward Elimination: Step 1

Variable stdx2x4 Removed: R-Square = 0.7417 and C(p) = 9.0036

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	9.82191	1.09132	137.16	<.0001
Error	430	3.42138	0.00796		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.79067	0.08419	0.70182	88.21	<.0001
graduation_rate	0.71724	0.04518	2.00535	252.03	<.0001
attendance_rate	0.78360	0.10632	0.43217	54.31	<.0001
pct_stu_safe	0.14152	0.06910	0.03338	4.20	0.0411
pct_stu_enough_variety	-0.01029	0.04352	0.00044522	0.06	0.8131
stdx1x2	0.02674	0.00671	0.12643	15.89	<.0001
stdx1x3	0.02532	0.00665	0.11536	14.50	0.0002
stdx1x4	-0.00483	0.00497	0.00751	0.94	0.3318
stdx2x3	-0.02317	0.00582	0.12593	15.83	<.0001
stdx3x4	0.00380	0.00454	0.00557	0.70	0.4034

In STEP 1, stdx_2x_4 has been deleted and as shown in the above table. Next, we have a high p-value (0.8131) for the predictor pct_stu_enogh_variety. Since 0.8131 is greater than 0.05, we eliminate it in the next step.

Table 23: Step 2 of Backward Elimination Method

Backward Elimination: Step 2

Variable pct_stu_enough_variety Removed: R-Square = 0.7416 and C(p) = 7.0594

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	9.82146	1.22768	154.63	<.0001
Error	431	3.42182	0.00794		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.79229	0.08381	0.70943	89.36	<.0001
graduation_rate	0.71792	0.04504	2.01752	254.12	<.0001
attendance_rate	0.78143	0.10581	0.43301	54.54	<.0001
pct_stu_safe	0.13566	0.06443	0.03520	4.43	0.0358
stdx1x2	0.02692	0.00666	0.12978	16.35	<.0001
stdx1x3	0.02517	0.00661	0.11503	14.49	0.0002
stdx1x4	-0.00498	0.00492	0.00814	1.03	0.3118
stdx2x3	-0.02294	0.00574	0.12695	15.99	<.0001
stdx3x4	0.00382	0.00453	0.00564	0.71	0.3997

In STEP 2, pct_stu_enogh_variety has been deleted and as shown above in the table results, we have the next highest p-value for the predictor stdx_3x_4. So, we eliminate it in the next step.

Table 24: Step 3 of Backward Elimination Method

Backward Elimination: Step 3

Variable stdx3x4 Removed: R-Square = 0.7412 and C(p) = 5.7670

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	9.81582	1.40226	176.74	<.0001
Error	432	3.42746	0.00793		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.77745	0.08192	0.71463	90.07	<.0001
graduation_rate	0.72434	0.04437	2.11422	266.48	<.0001
attendance_rate	0.76508	0.10398	0.42951	54.14	<.0001
pct_stu_safe	0.13049	0.06411	0.03287	4.14	0.0424
stdx1x2	0.02603	0.00657	0.12448	15.69	<.0001
stdx1x3	0.02630	0.00647	0.13103	16.52	<.0001
stdx1x4	-0.00264	0.00406	0.00336	0.42	0.5158
stdx2x3	-0.02306	0.00573	0.12837	16.18	<.0001

In STEP 3, stdx_3x_4 has been deleted and as shown above in the table results. The next highest p-value is for the predictor stdx_1x_4. After eliminating the stdx_3x_4, stdx_1x_4's p-value has gone up to 0.5158. So, we eliminate it in the next step.

Table 25: Step 4 of Backward Elimination Method

Backward Elimination: Step 4

Variable stdx1x4 Removed: R-Square = 0.7409 and C(p) = 4.1879

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	9.81246	1.63541	206.40	<.0001
Error	433	3.43082	0.00792		
Corrected Total	439	13.24328			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-0.78350	0.08133	0.73527	92.80	<.0001
graduation_rate	0.72538	0.04431	2.12300	267.94	<.0001
attendance_rate	0.77202	0.10336	0.44200	55.78	<.0001
pct_stu_safe	0.12942	0.06405	0.03235	4.08	0.0439
stdx1x2	0.02614	0.00657	0.12565	15.86	<.0001
stdx1x3	0.02576	0.00641	0.12779	16.13	<.0001
stdx2x3	-0.02350	0.00569	0.13521	17.06	<.0001

In STEP 4, predictor stdx_1x_4 has been eliminated. Our alpha value is 0.05. Therefore, we don't eliminate predictors furthermore. This provides us with the best model from this backward elimination method and it is the same as Model 6 from the best subset selection method.

We again conclude that the model with the above (table 25) six predictor variables is one of the potentially best models.

Table 26: Summary of Backward Elimination Method

Bounds on condition number: 2.5091, 76.623

All variables left in the model are significant at the 0.0500 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	stdx2x4	9	0.0000	0.7417	9.0036	0.00	0.9521
2	pct_stu_enough_variety	8	0.0000	0.7416	7.0594	0.06	0.8131
3	stdx3x4	7	0.0004	0.7412	5.7670	0.71	0.3997
4	stdx1x4	6	0.0003	0.7409	4.1879	0.42	0.5158

The above summary table shows the predictors that have been eliminated along with their r-square, C(p), F and P values.

In conclusion, based on the three Model Search Methods, we obtain two best models and they are model 6 and model 7:

Model 6:

$\widehat{College\_career\_rate}$ = -0.7835 + 0.72538*(graduation_rate) + 0.77202*(attendance_rate) + 0.12492*(pct_stu_safe) + 0.02614*(std x1 x2) + 0.02576* (std x1 x3) -0.0235*(std x2 x3)

Model 7:

$\widehat{College\_career\_rate}$ = -0.77745 + 0.72434*(graduation_rate) +0.76508*(attendance_rate) + 0.13049*(pct_stu_safe) + 0.02603*(stdx1x2) + 0.0263*(stdx1x3) -0.00264*(stdx1x4) -0.02306*(stdx2x3)

5. MODEL SELECTION

Model 7:

Model Form: $\widehat{College\_career\_rate}$ = -0.77745 + 0.72434*(graduation_rate) +0.76508*(attendance_rate) + 0.13049*(pct_stu_safe) + 0.02603*(stdx1x2) + 0.0263*(stdx1x3) -0.00264*(stdx1x4) -0.02306*(stdx2x3)

Table 27: ANOVA table for model 7

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Number of Observations Read	440
Number of Observations Used	440

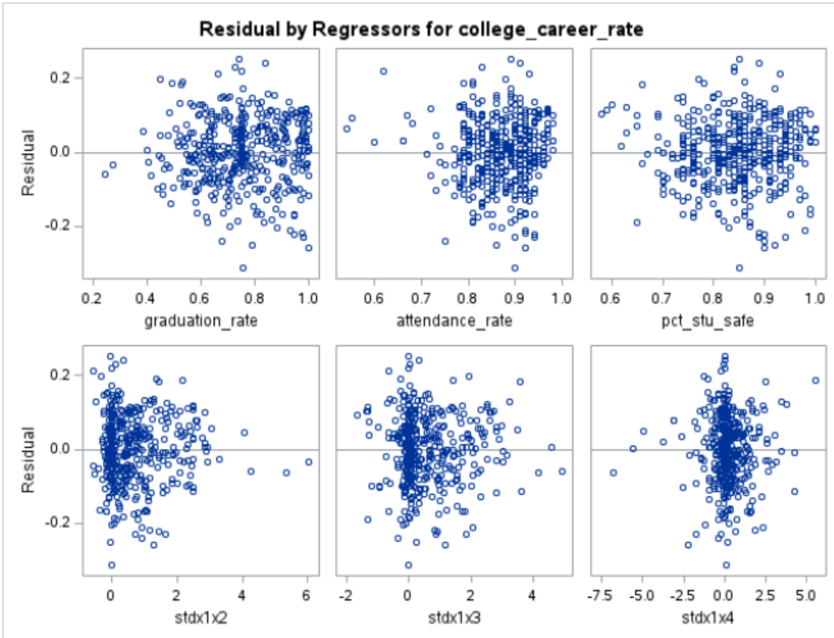
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	9.81582	1.40226	176.74	<.0001
Error	432	3.42746	0.00793		
Corrected Total	439	13.24328			

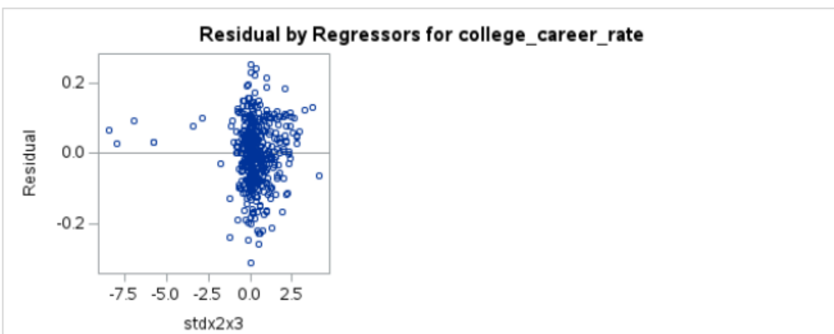
Root MSE	0.08907	R-Square	0.7412
Dependent Mean	0.56672	Adj R-Sq	0.7370
Coeff Var	15.71728		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.77745	0.08192	-9.49	<.0001	0
graduation_rate	1	0.72434	0.04437	16.32	<.0001	2.32322
attendance_rate	1	0.76508	0.10398	7.36	<.0001	2.53581
pct_stu_safe	1	0.13049	0.06411	2.04	0.0424	1.49490
stdx1x2	1	0.02603	0.00657	3.96	<.0001	1.97071
stdx1x3	1	0.02630	0.00647	4.06	<.0001	2.12304
stdx1x4	1	-0.00264	0.00406	-0.65	0.5158	1.09768
stdx2x3	1	-0.02306	0.00573	-4.02	<.0001	2.42464

Model 7 predicted **college_career_rate** based on **graduation_rate, attendance_rate, pct_stu_safe,** and interaction terms **stdx1x2, stdx1x3, stdx1x4,** and **stdx2x3.** The regression output indicated significant predictors, including graduation rate, attendance rate, and some interaction terms. The R-squared was 0.7412, and the adjusted R-squared was 0.7370, suggesting a good fit.

Figure 18: Residual vs. each predictor variables of model 7

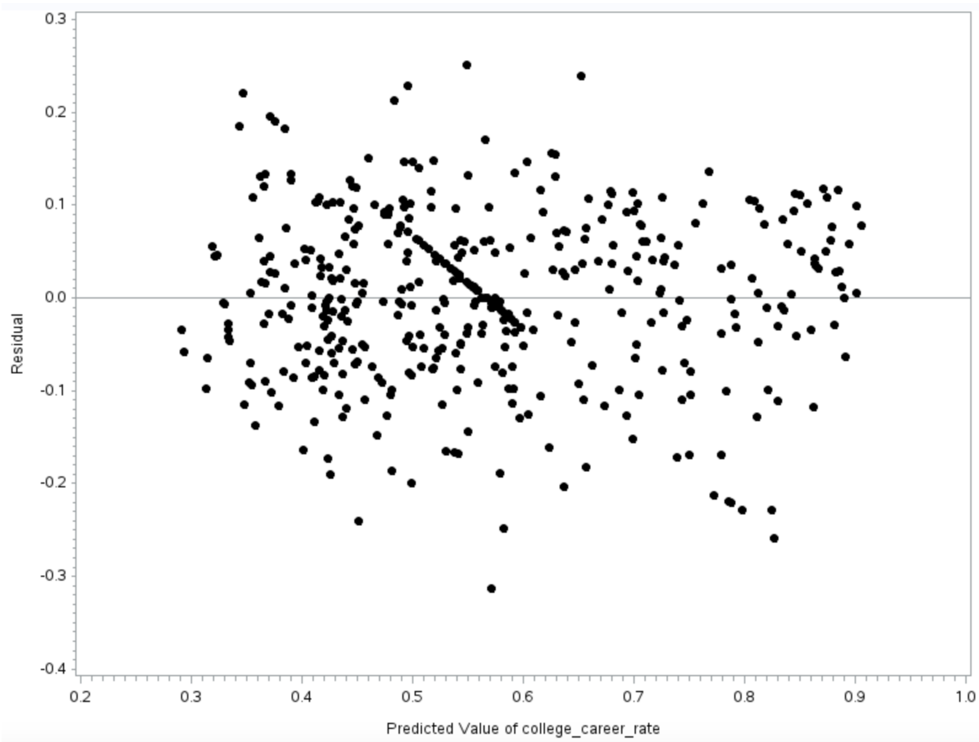




Assumption 1 : Model is reasonable

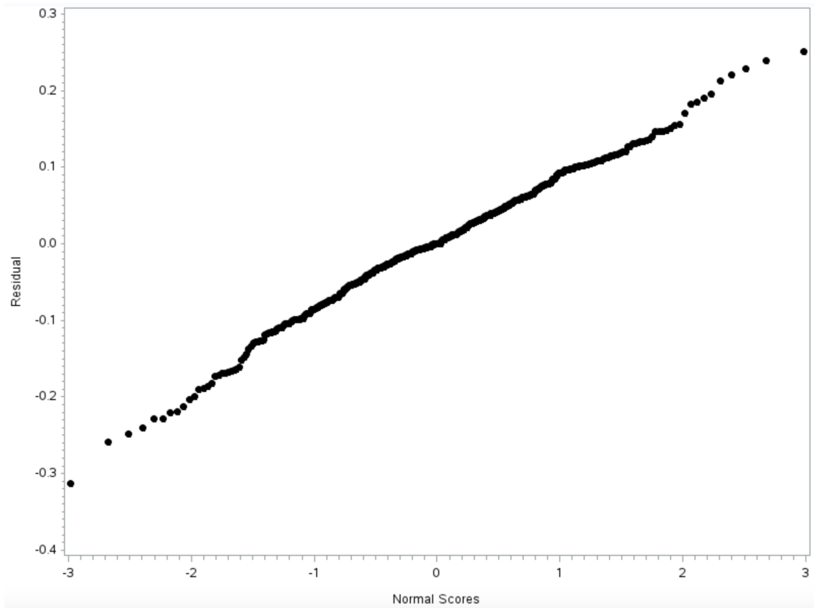
From Figure 18, we notice that there are random point clouds present and there is no curvature; therefore, the assumption that a linear model is appropriate is satisfied.

Figure 19: Residual Vs. Predicted college_career_rate for model 7



Assumption 2: Residuals have constant variance: From the residual vs Fitted values (Y_{hat}) plot in figure 20, we can notice that residuals seem to be randomly scattered across the horizontal line and no funnel shape exists; therefore, we can conclude that our assumption that the residuals have constant variance is satisfied. The random point clouds in the scale-location plot also confirm the assumption.

Figure 21: Normality Plot for Model 7



Assumption 3: Residuals are normally distributed: From Figure 21 we can see that points followed the normal line closely, but we notice some deviations at the ends of the plot, suggesting that extreme values might not be perfectly normally distributed.

Assumption 4: Predictors are not time related: Since our dataset wasn't collected in a timely manner, this assumption is accurate

Figure 23: Variance Inflation values for the predictors of model 7

```
VIF
# VIF of model 1
vif(model1)

## graduation_rate attendance_rate    pct_stu_safe         stdx1x2
stdx1x3
##        3.609232        2.570463        1.516527        2.268744
2.113112
##         stdx1x4         stdx2x3
##        3.175429        2.391364
```

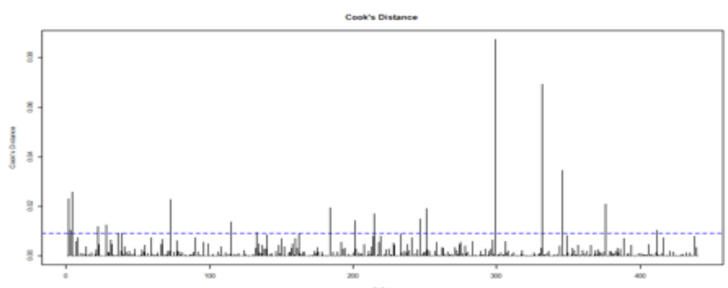
Assumption 5 : Residuals are uncorrelated :

From Figure 23, we notice that the VIF values for all the predictors are below 5. Hence, multicollinearity is not an issue.

Figure 24: Cook's Distance and Influential Outliers

```
# Outputing influential points and outliers beyond the cutoff based on Cook's
distance
cooks_distances <- cooks.distance(model1)
cutoff <- 4 / length(cooks_distances)
cooks_distances[cooks_distances > cutoff]

##          2          3          4         22         28         73
115
## 0.02305459 0.01032498 0.02591832 0.01175356 0.01241022 0.02285109
0.01375901
##        133        184        201        215        247        251
299
## 0.00960589 0.01957024 0.01429585 0.01695297 0.01499730 0.01916078
0.08728544
##        332        346        376        412
## 0.06930953 0.03449968 0.02102734 0.01045915
```



Influential Outliers (Cook's Distance): 18 points are identified as being beyond the cutoff F(0.5,8,432) which indicates that they have a disproportionately large influence on the model's coefficients. For example, observations indexed as 2, 4, and 73 show higher Cook's distance values, with the observation at index 346 showing the highest Cook's distance, suggesting it is particularly influential.

Model 7: F statistic & RSE values

```
## Residual standard error: 0.08836 on 432 degrees of freedom
## Multiple R-squared:  0.7453, Adjusted R-squared:  0.7412
## F-statistic: 180.6 on 7 and 432 DF,  p-value: < 2.2e-16
```

Model 6:

Model Form: $\widehat{College\_career\_rate}$ = -0.7835 + 0.72538*(graduation_rate) + 0.77202*(attendance_rate) + 0.12492*(pct_stu_safe) + 0.02614*(std x1 x2) + 0.02576* (std x1 x3) -0.0235*(std x2 x3)

Table 28: ANOVA Table for Model 6

The REG Procedure
Model: MODEL1
Dependent Variable: college_career_rate

Number of Observations Read	440
Number of Observations Used	440

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	9.81246	1.63541	206.40	<.0001
Error	433	3.43082	0.00792		
Corrected Total	439	13.24328			

Root MSE	0.08901	R-Square	0.7409
Dependent Mean	0.56672	Adj R-Sq	0.7373
Coeff Var	15.70681		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	-0.78350	0.08133	-9.63	<.0001	0
graduation_rate	1	0.72538	0.04431	16.37	<.0001	2.32023
attendance_rate	1	0.77202	0.10336	7.47	<.0001	2.50906
pct_stu_safe	1	0.12942	0.06405	2.02	0.0439	1.49391
stdx1x2	1	0.02614	0.00657	3.98	<.0001	1.96934
stdx1x3	1	0.02576	0.00641	4.02	<.0001	2.08710
stdx2x3	1	-0.02350	0.00569	-4.13	<.0001	2.39085

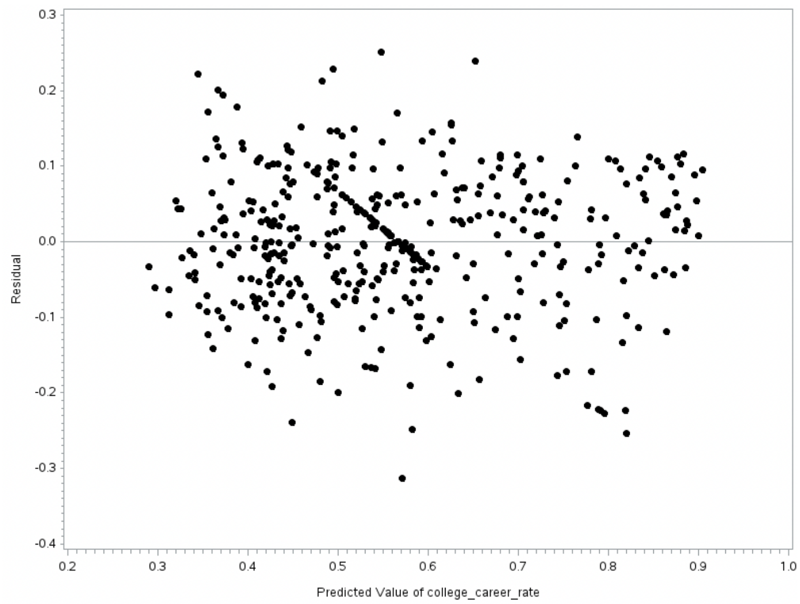
Model 6 predicted **college_career_rate** based on **graduation_rate, attendance_rate, pct_stu_safe,** and interaction terms **stdx1x2, stdx1x3,** and **stdx2x3.** The R-squared value was 0.7409, and the adjusted R-squared was 0.7373, suggesting a good fit.

Figure 25: Residual Vs. Each Predictor Variables of Model 6.

Assumption 1 : MLR model is reasonable

From Figure 25, we notice random point clouds with no visible curvature. This indicates that our assumption that an MLR model is reasonable is accurate.

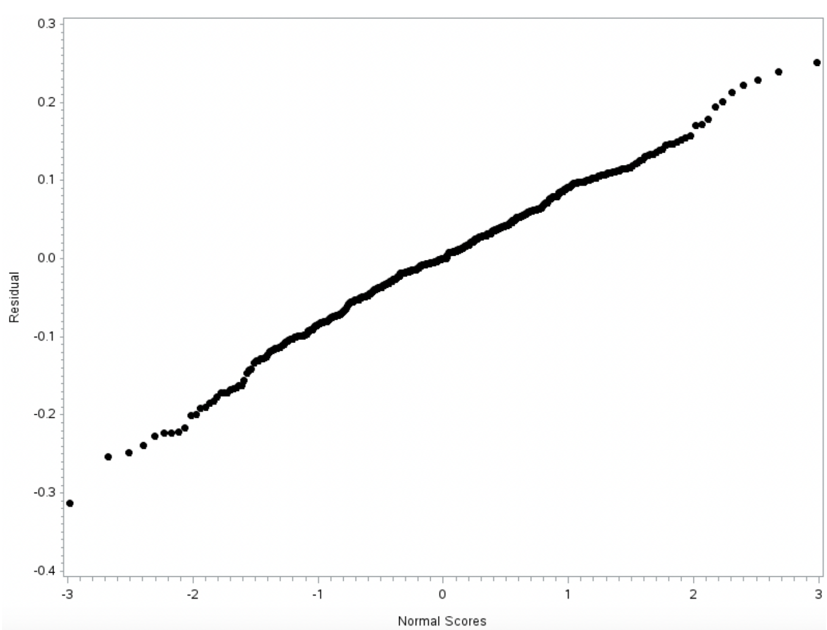
Figure 26: Residual Vs. Predicted college_career_rate for Model 6



Assumption 2: Residuals have constant variance.

From the Residuals vs predicted college_career_rate graph in Figure 26, we can notice that no funnel shape exists. This indicates that our assumption that the residuals have constant variance is accurate.

Figure 27: Normality Plot for Model 6



Assumption 3: Residuals are normally distributed.

Similar to Model 7 , the residuals are perfectly normally distributed, as can be seen in the normality plot in Figure 27. There is a mostly linear trend so normality is OK.

Assumption 4: Residuals are not time related.

Since our data is not collected in a timely manner, our assumption is satisfied

Figure 28: Variance Inflation Values for the variables of Model 6.

VIF

```
# VIF of model 2
vif(model2)

## graduation_rate attendance_rate    pct_stu_safe         stdx1x2
stdx1x3
##        2.320226        2.509059        1.493914        1.969339
2.087096
##         stdx2x3
##        2.390846
```

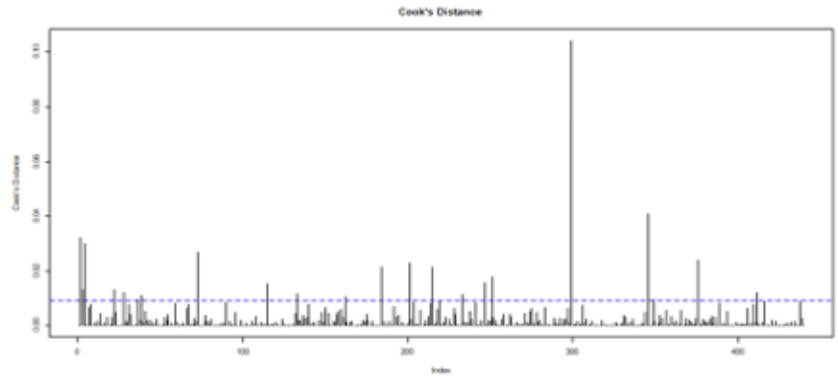
Assumption 5: Residuals are uncorrelated

All the VIF values are below 5, which suggests that multicollinearity is not a significant problem in this model.

Figure 29: Cook's Distance and Influential Outliers

```
# Ifluential points and outliers beyond the cutoff based on Cook's distance
cooks_distances <- cooks.distance(model2)
cutoff <- 4 / length(cooks_distances)
cooks_distances[cooks_distances > cutoff]

##           2           3           4          22          28          36
## 0.032111883 0.013271030 0.029861823 0.013290595 0.012192268 0.009685295
##          39          73         115         133         162         184
## 0.010813938 0.026818142 0.015321272 0.011647632 0.010455076 0.021264810
##         201         215         233         247         251         299
## 0.022720977 0.021312972 0.011221136 0.015570466 0.017876744 0.104115021
##         346         349         376         412
## 0.040820428 0.009485396 0.023808248 0.011975957
```



Influential Outliers (Cook's Distance):

Based on the cook's distance, we find 22 outliers. Although there are 22 outliers, value #299 seems to be the most highly influential outlier (way over the cutoff: F(0.5,7,433)). This indicates #299 warrants investigation and removal if necessary.

Model 6: F statistic value & RSE values

```
## Residual standard error: 0.08901 on 433 degrees of freedom
## Multiple R-squared:  0.7409, Adjusted R-squared:  0.7373
## F-statistic: 206.4 on 6 and 433 DF,  p-value: < 2.2e-16
```

Selection of the Best Model:

When selecting the ideal model between the 2, the following factors are considered: Adjusted R-square, residual standard error, significance of the predictors, F-statistic, Cook's distance, and VIF

Model 7 has a higher adjR-squared of 0.7412, with Model 6 having a closer value of 0.7373. Model 7 has a RSE = 0.0884, while Model 6 has a RSE = 0.08901. Again, the values are very close, with Model 7 having a slightly lower RSE, which indicates a marginally better fit.

Model 7 has a slightly lower F-statistic of 180.6 as compared to Model 6 with 206.4. A higher F-statistic would indicate that the model is statistically significant. Also, all predictors in both models are statistically significant at the 0.05 level. In Model 7, one predictor has p-values greater than 0.05, indicating that it is not statistically significant at the 5% level. However, Model 6 has all predictors being significant at the 0.05 level.

The variance inflation factor indicates that Model 7 has all VIF values that are less than the threshold of 5. Model 6 has VIF values lower than 5 and also lower than those of model 7, indicating that it has less multicollinearity, which is preferable for the stability of the model coefficients. Points with a high Cook's distance are considered to be influential to the regression model. Both models have influential points.

Model 6 is preferable because it maintains a similar explanatory power to Model 7 (as indicated by the comparable adjusted R-squared values) but has improved in terms of statistical significance (higher F-statistic), has all significant predictors, and exhibits reduced multicollinearity (lower VIF values). The choice of Model 6 also aligns with the principle of parsimony, which favors simpler models when they perform almost as well as more complex ones.

6. Final MLR Model

So, our overall best model is:

$College\widehat{\_career\_rate}$ = -0.7835 + 0.72538*(graduation_rate) + 0.77202*(attendance_rate) + 0.12492*(pct_stu_safe) + 0.02614*(std x1 x2) + 0.02576* (std x1 x3) -0.0235*(std x2 x3)

The model obtained an R-Square value of 0.7373, so the predictor variables are linearly associated with college_career_rate.

The stdx_1x_2 is the standardized interaction term between graduation_rate and attendance_rate. The std x_1x_3 is the standardized interaction term between graduation_rate and pct_stu_safe. The stdx_2x_3 is the standardized interaction term between attendance_rate and pct_stu_safe

Joint C.I for the Parameters:

Estimating β_0, β_1, β_2, β_3, β_4, β_5, β_6 simultaneously with a 95% family confidence coefficient.

β_0, β_1, β_2, β_3, β_4, β_5, β_6 are the k^{th} diagonal element in the table.

Table 29:

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Variance Inflation
Intercept	1	-0.78350	0.08133	-9.63	<.0001	141.31457	0
graduation_rate	1	0.72538	0.04431	16.37	<.0001	8.90746	2.32023
attendance_rate	1	0.77202	0.10336	7.47	<.0001	0.34237	2.50906
pct_stu_safe	1	0.12942	0.06405	2.02	0.0439	0.08863	1.49391
stdx1x2	1	0.02614	0.00657	3.98	<.0001	0.28504	1.96934
stdx1x3	1	0.02576	0.00641	4.02	<.0001	0.05375	2.08710
stdx2x3	1	-0.02350	0.00569	-4.13	<.0001	0.13521	2.39085

From the Anova table, the standard errors for parameters are as following:

$s\{b_0\} = 0.08133$

$s\{b_1\} = 0.04431$

$s\{b_2\} = 0.10336$

$s\{b_3\} = 0.06405$

$s\{b_4\} = 0.00657$

$s\{b_5\} = 0.00641$

$s\{b_6\} = 0.00569$

Simultaneous Two-sided 95% joint confidence interval for $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, and\ \beta_6$:

$b_k \pm B\, s\{b_k\}$, where

$B = t(1 - \alpha/2g; n - p)$

$= t(1- 0.05 /2*6; 440-7)$

$= t(1- 0.0041667; 433)$

$=t(0.9958; 433) = 2.5798$

$\beta_0 = b_0 \pm B\, s\{b0\}$ $= -0.78350 \pm (2.5798 * 0.08133)$

$= (-0.9933 , -0.5737)$

$\beta_1 = b_1 \pm$ B s{b1} $= 0.72538 \pm (2.5798 * 0.04431)$

$= (0.6111 , 0.8397)$

$\beta_2 = b_2 \pm$B s{b2} $= 0.77202 \pm (2.5798 * 0.10336)$

$= (0.5054 , 1.0387)$

$\beta_3 = b_3 \pm$B s{b3} $= 0.12942 \pm (2.5798 * 0.06405)$

$= (-0.0358, 0.2947)$

$\beta_4 = b_4 \pm$B s{b4} $= 0.02614 \pm (2.5798 * 0.00657)$

$= (0.0092, 0.0431)$

$\beta_5 = b_5 \pm$B s{b5} $= 0.02576 \pm (2.5798 * 0.00641)$

$= (0.0092, 0.0423)$

$\beta_6 = b_6 \pm$B s{b6} $= -0.0235 \pm (2.5798 * 0.00569)$

$= (-0.0382, -0.0088)$

We are 95% confident that β_0 is contained in (-0.9933 , -0.5737), β_1 is contained in (0.6111 , 0.8397), $\beta_2$is contained in (0.5054 , 1.0387), β_3 is contained in (-0.0358, 0.2947), β_4 is contained in (0.0092, 0.0431), β_5 is contained in (0.0092, 0.0423), and β_6 is contained in (-0.0382, -0.0088) simultaneously.

Confidence Interval for X_{hnew} :

For X_{hnew} we consider graduation_rate = 0.6, attendance_rate= 0.7, pct_stu_safe= 0.8.

To calculate the standardized interaction terms, we need the mean and standard deviation of graduation_rate, attendance_rate, and pct_stu_safe.

Variable	N	Mean	Std Dev
college_career_rate	440	0.56672	0.17369
graduation_rate	440	0.75244	0.14603
attendance_rate	440	0.86998	0.06510
pct_stu_safe	440	0.84492	0.08107

$\text{std}x_1x_2 = \text{std}x_1 * \text{std}x_2 = [\ (0.6-0.75244) / 0.14603] * [\ (0.7-0.86998) / 0.06510]$

$= -1.0439 * -2.6111 = 2.7257$

$\text{std}x_1x_3 = \text{std}x_1 * \text{std}x_3 = [\ (0.6-0.75244) / 0.14603] * [\ (0.8-0.84492) / 0.08107]$

= -1.0439 * -0.5541 = 0.5784

stdx_2x_3 = stdx_2 * stdx_3 = [(0.7-0.86998) / 0.06510] * [(0.8-0.84492) / 0.08107]

= -2.6111 * -0.5541 = 1.4468

X_{hnew}^T = [1 0.6 0.7 0.8 2.7257 0.5784 1.4468]

$\widehat{Y}_h$ = -0.7835*(1) + 0.72538* (0.6) + 0.77202 * (0.7) + 0.12942* (0.8) + 0.02614* (2.7257)+ 0.02576* (0.5784) -0.02350* (1.4468)

$\widehat{Y}_h$ = 0.3479

Standard error Calculation:

$s^2\{\widehat{Y}_h\}$ = MSE* (X_{hnew}^T $(X^TX)^{-1}$ X_{hnew})

By calculating (X_{hnew}^T $(X^TX)^{-1}$ X_{hnew}) on excel, we obtain the value as 0.050038

Therefore,

$s^2\{\widehat{Y}_h\}$ = 0.00792 * 0.050038 = 0.000396

s$\{\widehat{Y}_h\}$ = $\sqrt{0.000396}$ = 0.0199

Using this value, We find C. I's

C.I. = $\widehat{Y}_h$ ± t(1-α/2g,n-p)*s$\{\widehat{Y}_h\}$ = 0.3479 ± 1.960*0.0199 = (0.3089 , 0.3869)

Therefore, we are 95 % confident that the college_career_rate for the X_{hnew} will lie between 0.3089 and 0.3869.

Confidence Band Calculations:

Confidence bands are calculated using the formula:

CB = $\widehat{Y}_h$ ± W*s$\{\widehat{Y}_h\}$

W value is obtained from the equation: W^2 = p* F(1-α,p,n-p)

$W^2 = 7 * F(0.95, 7, 433) = 7 * 2.03073$

We obtain W=3.77

Therefore, **CB**= 0.3479 ± 3.77*0.0199 = **(0.2729, 0.4229)**

Hence, we are 95 % confident that true line for college_career_rate will lie between 0.2729 and 0.4229. It is observable that the confidence band is a lot wider than the confidence interval.

Prediction Interval calculation:

We calculate the prediction interval using the following formula:

$$P.I. = \widehat{Y}_h \pm t(1-\alpha/2, n-p) * S\{pred\}$$

Here, $S\{pred\} = \sqrt{s^2\{\widehat{Y}_h\} + MSE}$

We know, MSE= 0.00792 and $s^2\{\widehat{Y}_h\} = 0.000396$

S{Pred} = $\sqrt{0.000396 + 0.00792} = 0.0912$

Hence, **P.I.** = 0.3479 ± 1.960*0.0912 = 0.3479 ± 0.1788 = **(0.1691,0.5267)**

Therefore, we are 95% confident that the mean college_career_rate for new future observations will lie between **0.1691 and 0.5267**. We also observe that the Prediction interval is wider than C.I. and C.B and this is because of the uncertainty when predicting for new future observations.

7. Final Discussion:

In conclusion, our model is complete and verified. A 7-parameter model for predicting future values of College_career_rate was selected. The final model was obtained by performing a model search for the two best models and testing these models for model assumptions (normality, constant variance, and multicollinearity) and influential outliers. The final model contains the predictor variables attendance_rate, graduation_rate ,pct_stu_safe, and the standardized interaction terms std x1 x2, std x1 x3, std x2 x3. We can start using this model after using it for a test dataset.

As we add more predictor variables to the model, we must verify the model assumptions, check for serious multicollinearity, and follow the above steps again. After adding new predictor variables, we may find that the variance is non-constant. In such cases, we can use a Weighted

Least Squares regression analysis. Furthermore, we can also use the Robust Regression to reduce the influence of outliers. This model can be applied in the future to mathematically predict the prospects for future students at different schools. This can help students plan their applications and select their area of study without fear of choosing a school that doesn't meet their personal requirements.