Multiple Linear Regression Analysis

Stat 481

Erica Castillo

Summary:

The objective in this project is to analyze specific characteristics in a CPS dataset that influences college enrollment for students. The goal is to build a predictive model that highlights the strongest indicators for college enrollment. By performing a multiple linear regression analysis, I will be able to see what factors influence college-bound individuals the most.

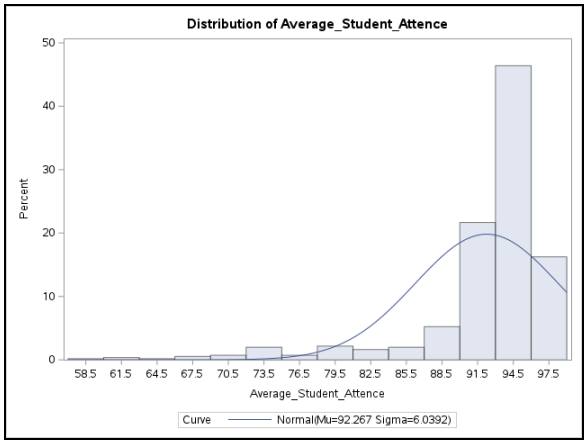
Data Introduction and Description

The CPS dataset (CPS_ES-HS_ProgressReport_2011-2012.csv.), labeled as "CPS_DATA," has 25 variables. In this project, we will specifically look at 9 variables and how they affect the variable College_Enrollment (our response variable). The 9 variables (the predictor variables) of interest are:

1. Average_Student_Attence (X1)
2. Rate_of_Misconducts (X2)
3. Average_Teacher_Attence (X3)
4. X9_Grade_Explore_2009 (X4)
5. X11_Grade_Average_ACT_2011 (X5)
6. College_Eligibility (X6)
7. Graduation_Rate (X7)
8. Freshman_on_Track_Rate (X8)
9. Probation (X9)

Descriptive Statistics

Before any analysis is performed, I need to see what our data looks like. Below are some summary statistics for each of the variables we are working with.

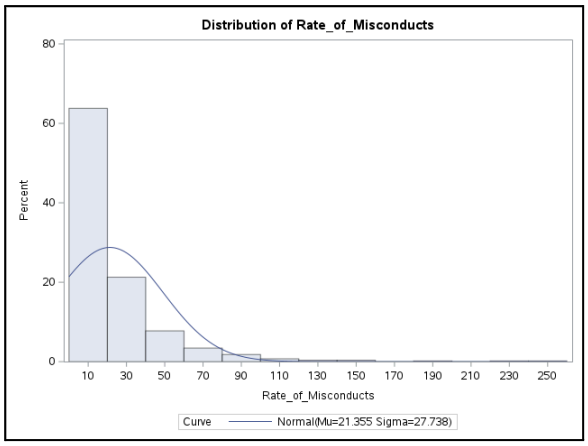


The UNIVARIATE Procedure
Variable: Average_Student_Attence

Moments			
N	554	Sum Weights	554
Mean	92.2669675	Sum Observations	51115.9
Std Deviation	6.03919469	Variance	36.4718725
Skewness	-2.7351076	Kurtosis	8.25728129
Uncorrected SS	4736478.03	Corrected SS	20168.9455
Coeff Variation	6.54534863	Std Error Mean	0.2565808

Basic Statistical Measures			
Location		Variability	
Mean	92.26697	Std Deviation	6.03919
Median	94.35000	Variance	36.47187
Mode	95.50000	Range	40.50000
		Interquartile Range	3.80000

The Average_Student_Attence histogram is skewed to the left, indicating it is negatively skewed. This means that the variable's mean is less than its median. The “Basic Statistical Measures” chart confirms the negative skewness. It also shows that the sample size for this variable is 554; Which means that the data is taken from 554 individuals.

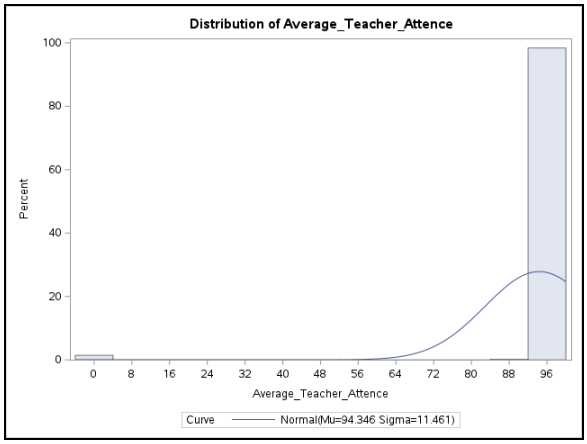


The UNIVARIATE Procedure
Variable: Rate_of_Misconducts

Moments			
N	555	Sum Weights	555
Mean	21.3554955	Sum Observations	11852.3
Std Deviation	27.7380836	Variance	769.401283
Skewness	3.50716303	Kurtosis	18.7728022
Uncorrected SS	679360.05	Corrected SS	426248.311
Coeff Variation	129.887333	Std Error Mean	1.17741608

Basic Statistical Measures			
Location		Variability	
Mean	21.35550	Std Deviation	27.73808
Median	12.40000	Variance	769.40128
Mode	0.00000	Range	251.60000
		Interquartile Range	22.90000

The Rate_of_Misconducts histogram is skewed to the right, indicating it is positively skewed. This means that the variable's mean is greater than its median. The “Basic Statistical Measures” chart confirms the positive skewness. It also shows that the sample size for this variable is 555; Which means that the data is taken from 555 individuals.

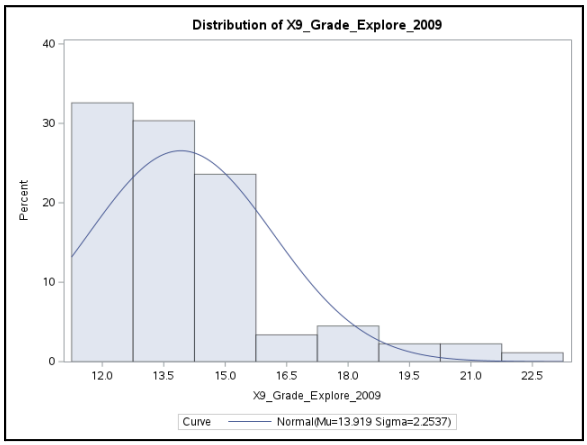


The UNIVARIATE Procedure
Variable: Average_Teacher_Attence

Moments			
N	555	Sum Weights	555
Mean	94.3464865	Sum Observations	52362.3
Std Deviation	11.4614359	Variance	131.364514
Skewness	-8.079531	Kurtosis	64.0086444
Uncorrected SS	5012974.97	Corrected SS	72775.9406
Coeff Variation	12.1482382	Std Error Mean	0.48651086

Basic Statistical Measures			
Location		Variability	
Mean	94.34649	Std Deviation	11.46144
Median	95.80000	Variance	131.36451
Mode	95.90000	Range	98.50000
		Interquartile Range	1.30000

The Average_Teacher_Attence histogram is skewed to the left, indicating it is negatively skewed. This means that the variable's mean is less than its median. The “Basic Statistical Measures” chart confirms the negative skewness. It also shows that the sample size for this variable is 555; Which means that the data is taken from 555 individuals.

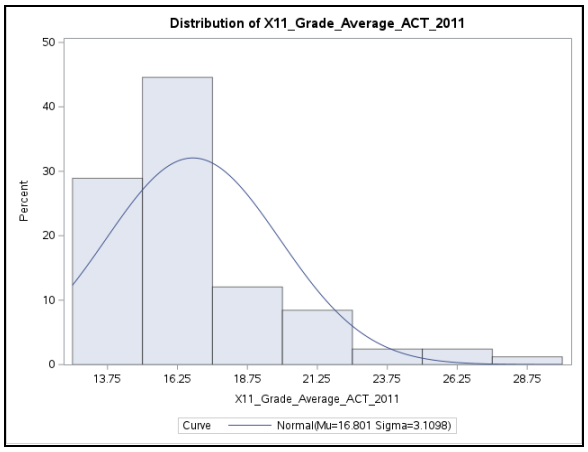


The UNIVARIATE Procedure
Variable: X9_Grade_Explore_2009

Moments			
N	89	Sum Weights	89
Mean	13.9191011	Sum Observations	1238.8
Std Deviation	2.25367762	Variance	5.07906282
Skewness	1.71373096	Kurtosis	3.24295525
Uncorrected SS	17689.94	Corrected SS	446.957528
Coeff Variation	16.1912583	Std Error Mean	0.23888935

Basic Statistical Measures			
Location		Variability	
Mean	13.91910	Std Deviation	2.25368
Median	13.20000	Variance	5.07906
Mode	12.00000	Range	10.90000
		Interquartile Range	2.30000

The X9_Grade_Explore_2009 histogram is skewed to the right, indicating it is positively skewed. This means that the variable's mean is greater than its median. The “Basic Statistical Measures” chart confirms the positive skewness. It also shows that the sample size for this variable is 89; Which means that the data is taken from 89 individuals.

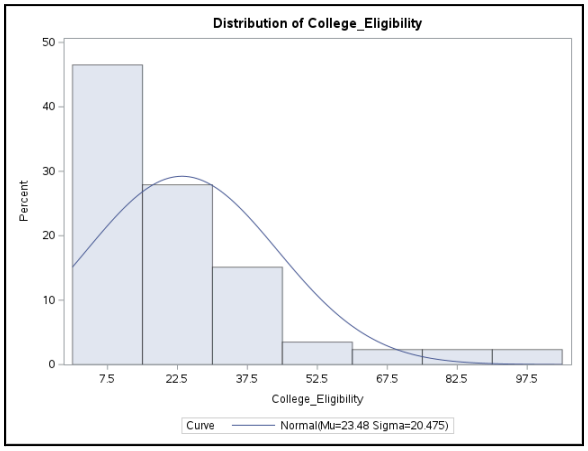


The UNIVARIATE Procedure
Variable: X11_Grade_Average_ACT_2011

Moments			
N	83	Sum Weights	83
Mean	16.8012048	Sum Observations	1394.5
Std Deviation	3.10976017	Variance	9.67060829
Skewness	1.95034871	Kurtosis	3.97930255
Uncorrected SS	24222.27	Corrected SS	792.98988
Coeff Variation	18.5091498	Std Error Mean	0.34134052

Basic Statistical Measures			
Location		Variability	
Mean	16.80120	Std Deviation	3.10976
Median	15.70000	Variance	9.67061
Mode	14.00000	Range	15.30000
		Interquartile Range	2.90000

The X11_Grade_Average_ACT_2011 histogram is skewed to the right, indicating it is positively skewed. This means that the variable's mean is greater than its median. The “Basic Statistical Measures” chart confirms the positive skewness. It also shows that the sample size for this variable is 83; Which means that the data is taken from 83 individuals.

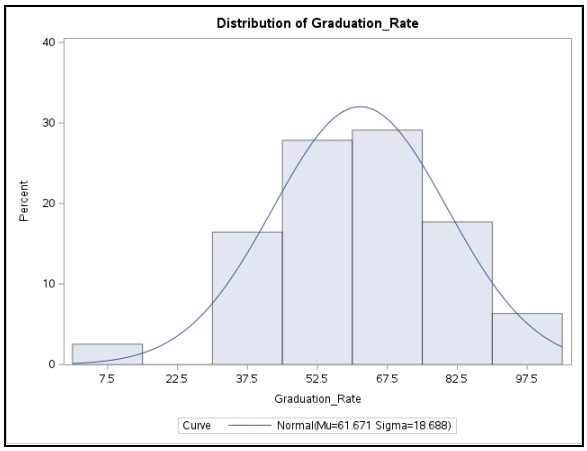


The UNIVARIATE Procedure
Variable: College_Eligibility

Moments			
N	86	Sum Weights	86
Mean	23.4802326	Sum Observations	2019.3
Std Deviation	20.4748961	Variance	419.221369
Skewness	2.01947832	Kurtosis	4.26525527
Uncorrected SS	83047.45	Corrected SS	35633.8164
Coeff Variation	87.2005677	Std Error Mean	2.20786483

Basic Statistical Measures			
Location		Variability	
Mean	23.48023	Std Deviation	20.47490
Median	16.50000	Variance	419.22137
Mode	14.80000	Range	95.50000
		Interquartile Range	19.30000

The College_Eligibility histogram is skewed to the right, indicating it is positively skewed. This means that the variable's mean is greater than its median. The “Basic Statistical Measures” chart confirms the positive skewness. It also shows that the sample size for this variable is 86; Which means that the data is taken from 86 individuals.

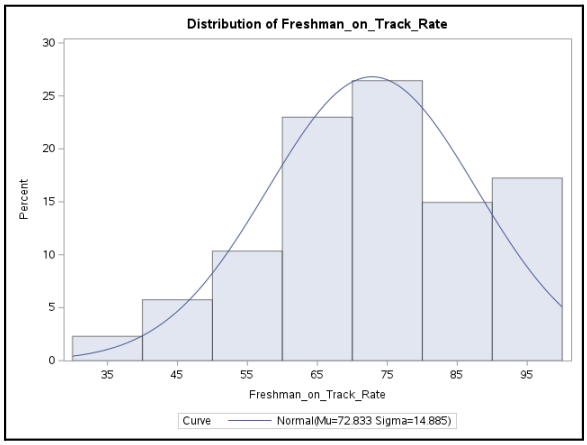


The UNIVARIATE Procedure
Variable: Graduation_Rate

Moments			
N	79	Sum Weights	79
Mean	61.6708861	Sum Observations	4872
Std Deviation	18.6876289	Variance	349.227475
Skewness	-0.4849541	Kurtosis	0.78411848
Uncorrected SS	327700.3	Corrected SS	27239.743
Coeff Variation	30.3021898	Std Error Mean	2.10252252

Basic Statistical Measures			
Location		Variability	
Mean	61.67089	Std Deviation	18.68763
Median	62.10000	Variance	349.22747
Mode	49.50000	Range	97.60000
		Interquartile Range	25.30000

The Graduation_Rate histogram is bell shaped indicating a normal distribution. This means that the variable's mean is approximately the same as its median. The “Basic Statistical Measures” chart confirms it is normally distributed. It also shows that the sample size for this variable is 79; Which means that the data is taken from 79 individuals.

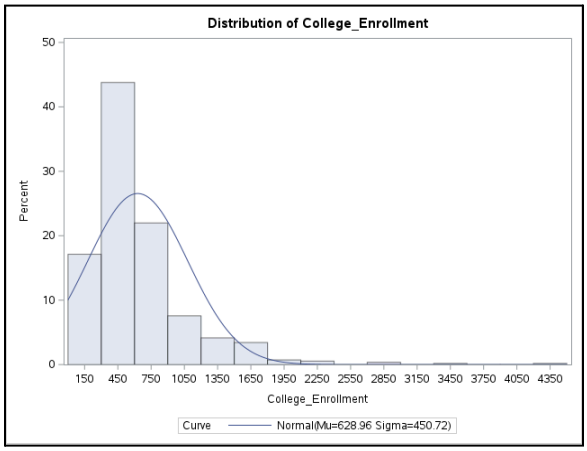


The UNIVARIATE Procedure
Variable: Freshman_on_Track_Rate

Moments			
N	87	Sum Weights	87
Mean	72.8333333	Sum Observations	6336.5
Std Deviation	14.8850844	Variance	221.565736
Skewness	-0.4776801	Kurtosis	-0.1713321
Uncorrected SS	480563.07	Corrected SS	19054.6533
Coeff Variation	20.4371868	Std Error Mean	1.59584855

Basic Statistical Measures			
Location		Variability	
Mean	72.83333	Std Deviation	14.88508
Median	74.10000	Variance	221.56574
Mode	90.70000	Range	66.00000
		Interquartile Range	23.00000

The Track_Rate histogram is bell shaped indicating a normal distribution. This means that the variable's mean is approximately the same as its median. The “Basic Statistical Measures” chart confirms it is normally distributed. It also shows that the sample size for this variable is 87; Which means that the data is taken from 87 individuals.



The UNIVARIATE Procedure
Variable: College_Enrollment

Moments			
N	555	Sum Weights	555
Mean	628.96036	Sum Observations	349073
Std Deviation	450.724161	Variance	203152.269
Skewness	2.70779417	Kurtosis	13.0352498
Uncorrected SS	332099437	Corrected SS	112546357
Coeff Variation	71.6617754	Std Error Mean	19.1321751

Basic Statistical Measures			
Location		Variability	
Mean	628.9604	Std Deviation	450.72416
Median	505.0000	Variance	203152
Mode	363.0000	Range	4347
		Interquartile Range	450.00000

The College_Enrollment histogram is skewed to the right, indicating it is positively skewed. This means that the variable's mean is greater than its median. The “Basic Statistical Measures” chart confirms the positive skewness. It also shows that the sample size for this variable is 555; Which means that the data is taken from 555 individuals.

The MEANS Procedure

Variable	N	N Miss
Average_Student_Attence	554	1
Rate_of_Misconducts	555	0
Average_Teacher_Attence	555	0
X9_Grade_Explore_2009	89	466
X11_Grade_Average_ACT_2011	83	472
College_Eligibility	86	469
Graduation_Rate	79	476
Freshman_on_Track_Rate	87	468
College_Enrollment	555	0
Probation	555	0

Based on the Means Procedure chart, we can see the sample size in each of our variables and how many missing values they have. For this project, I will leave the missing values in the model.

The FREQ Procedure

Probation	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	312	56.22	312	56.22
1	243	43.78	555	100.00

The variable Procedure is an indicator variable. This means it can only take two values. In this instance, 0 or 1. The chart shows the amount for the two values.

Multiple Linear Regression Analysis

During this analysis, I will use alpha value $\alpha = 0.05$ to make my decisions.

Multicollinearity Check

In this section, I will check for multicollinearity in our variables. In other words, if any of our explanatory variables have similar information and are repetitive, they need to be removed from the data. I will use the Variance Inflation Factor (VIF) technique to see which variables need to be removed. Any values greater than 10 will be removed from the data one by one.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4062.65900	12539	0.32	0.7470	0
Average_Student_Attence	1	30.56895	19.81489	1.54	0.1278	4.41298
Rate_of_Misconducts	1	-11.94121	8.36826	-1.43	0.1585	1.87547
Average_Teacher_Attence	1	-46.86559	134.67129	-0.35	0.7290	1.58371
X9_Grade_Explore_2009	1	176.24490	193.70534	0.91	0.3663	29.28910
X11_Grade_Average_ACT_2011	1	-110.56721	149.95137	-0.74	0.4636	32.78001
College_Eligibility	1	11.71685	13.92915	0.84	0.4034	11.58742
Graduation_Rate	1	2.92289	9.39008	0.31	0.7566	3.50290
Freshman_on_Track_Rate	1	-28.24569	9.95584	-2.84	0.0061	2.96282
Probation	1	287.95536	290.11185	0.99	0.3247	2.83266

In this chart, the column of interest is the Variance Inflation column. As previously stated, a value greater than 10 needs to be removed from the data. Already I see that the variables X9, X11, and College_Eligibility are greater than 10, indicating multicollinearity. For this instance, we will removed X11 since it has the greater VIF value and then rerun the model.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	4977.07670	12434	0.40	0.6903	0
Average_Student_Attence	1	29.65289	19.70637	1.50	0.1372	4.39563
Rate_of_Misconducts	1	-12.13981	8.33452	-1.46	0.1500	1.87352
Average_Teacher_Attence	1	-59.68713	133.07440	-0.45	0.6553	1.55730
X9_Grade_Explore_2009	1	66.19765	123.04590	0.54	0.5924	11.90191
College_Eligibility	1	7.03564	12.35463	0.57	0.5710	9.18029
Graduation_Rate	1	3.96483	9.25050	0.43	0.6696	3.42357
Freshman_on_Track_Rate	1	-27.09531	9.79827	-2.77	0.0074	2.89006
Probation	1	333.66069	282.41599	1.18	0.2417	2.70335

After rerunning the model, it appears that X9's VIF value is still greater than 10. In this instance, we will remove the X9 variable from the model and rerun.

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	6342.95614	12106	0.52	0.6021	0
Average_Student_Attence	1	29.77727	19.59865	1.52	0.1334	4.39503
Rate_of_Misconducts	1	-13.18443	8.06143	-1.64	0.1067	1.77184
Average_Teacher_Attence	1	-66.55353	131.74597	-0.51	0.6151	1.54298
College_Eligibility	1	12.68118	6.48544	1.96	0.0548	2.55727
Graduation_Rate	1	5.31232	8.85690	0.60	0.5507	3.17259
Freshman_on_Track_Rate	1	-26.74916	9.72434	-2.75	0.0077	2.87760
Probation	1	319.57577	279.68206	1.14	0.2573	2.68012

After removing the variables X11 and X9, the remaining variables have a VIF value less than 10 indicating there is no multicollinearity present in the model.

Initial Regression Model

I will now perform the initial multiple regression analysis on our current model. The purpose is to see the relationship between college enrollment and our remaining predictor values.

The statistical model we currently have is:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \beta 3X3 + \beta 6X6 + \beta 7X7 + \beta 8X8 + \beta 9X9 + \epsilon$$

Or

$Y = \beta 0 + \beta 1(Average\_Student\_Attence) + \beta 2(Rate\_of\_Misconducts) +$

$\beta 3(Average\_Teacher\_Attence) + \beta 6(College\_Eligibility) + \beta 7(Graduation\_Rate) +$

$\beta 8(Freshman\_on\_Track\_Rate) + \beta 9(Probation)$

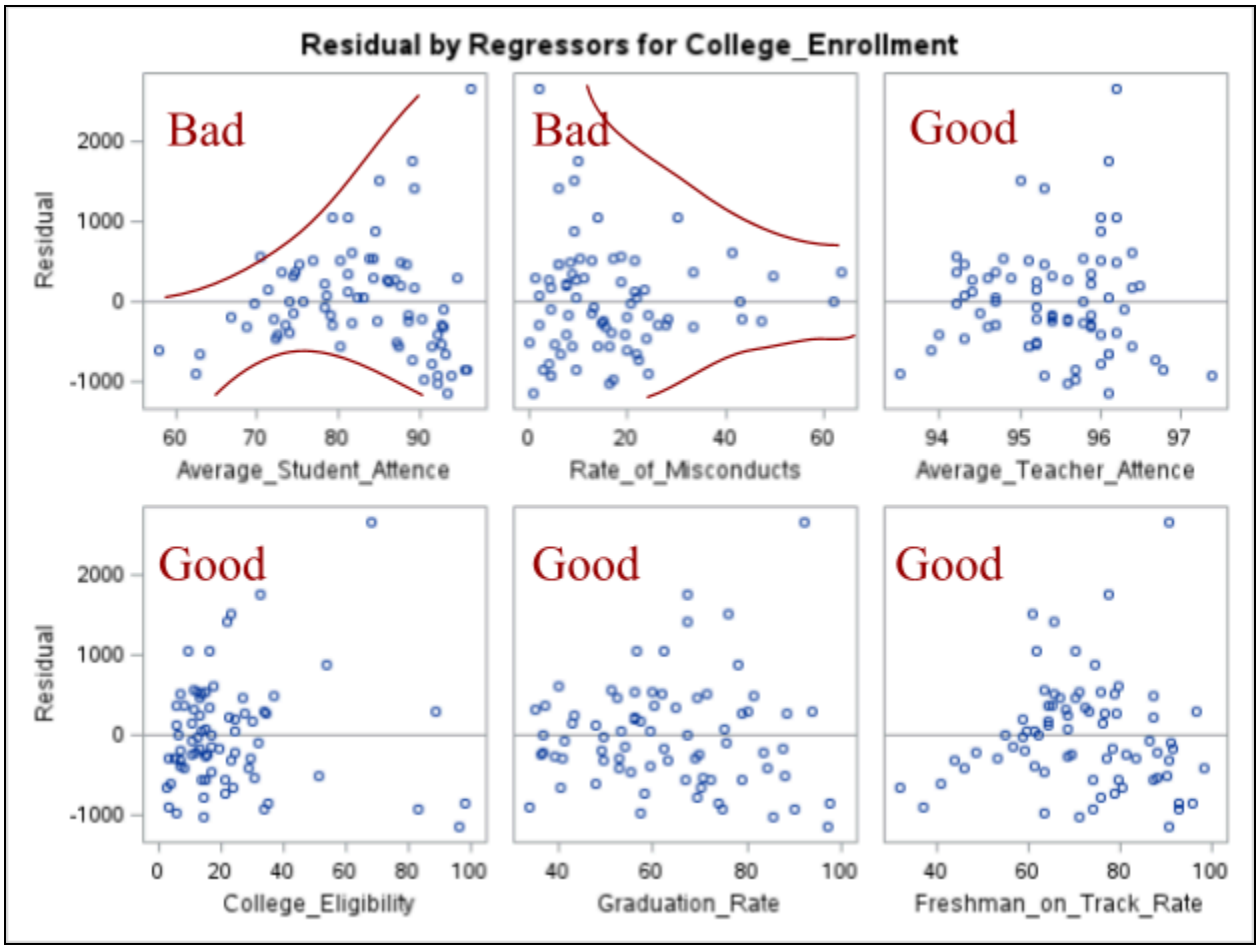
Checking Model Assumptions

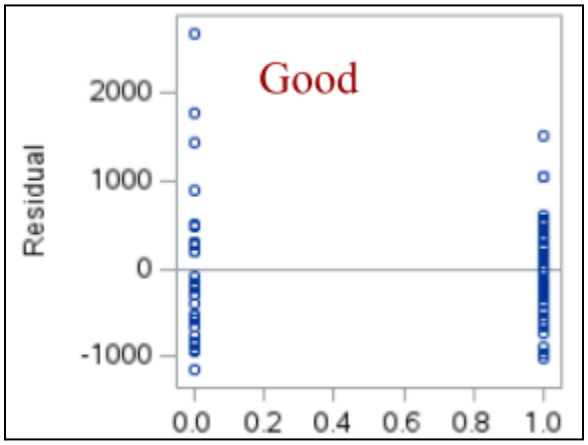
Basic model assumption are:

- E(Y*i*) at each value of X*i* is a linear function of X*i*
- Errors Ɛ*i* are independent
- Errors Ɛ*i* at each value of X*i* are normally distributed
- Errors Ɛ*i* at each value of X*i* have equal variances

<u>We do not need to do an independence check since that is only checked for time series data.</u>

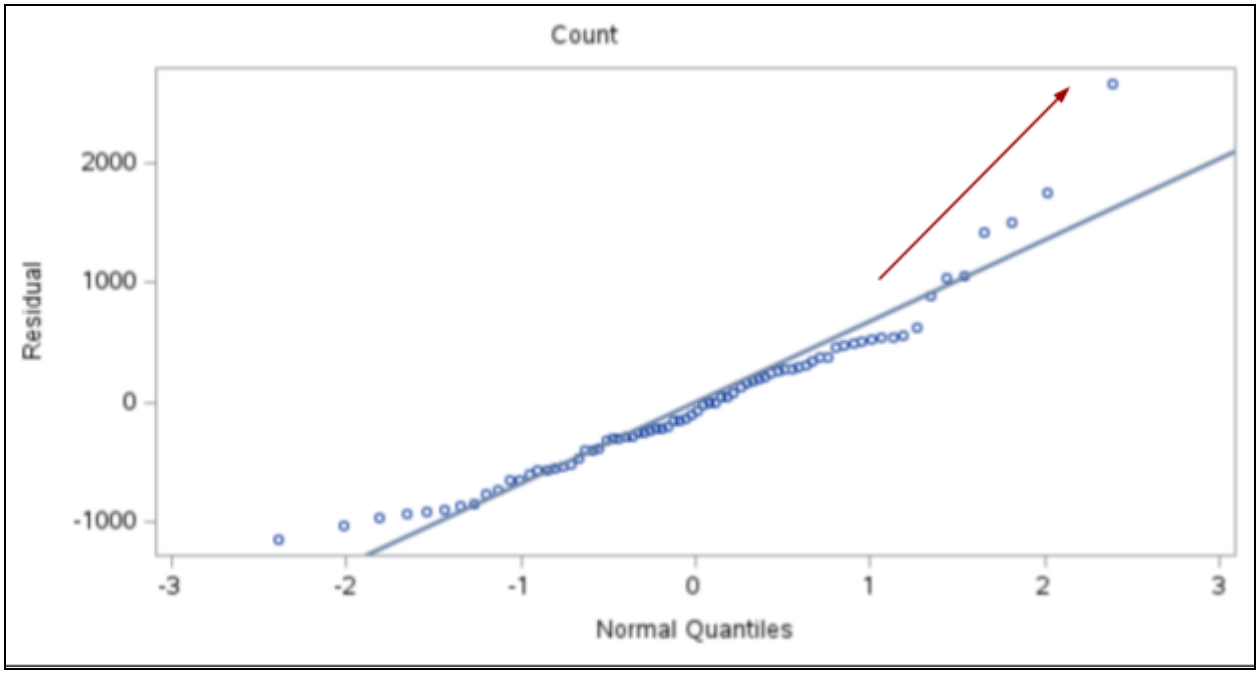
Check for Linearity





Based on the X*i* versus residual plot, there is a linearity violation. We want to see a random spread for the points. There should be no pattern present in the plots. Average_Student_Attence and Rate_of_Misconducts appear to have a fanned pattern which is not good.

Check for Normality



To test for normality, I check the residuals versus normal quantiles plot. The points start to deviate from the line which is indicating the residuals are not normal.

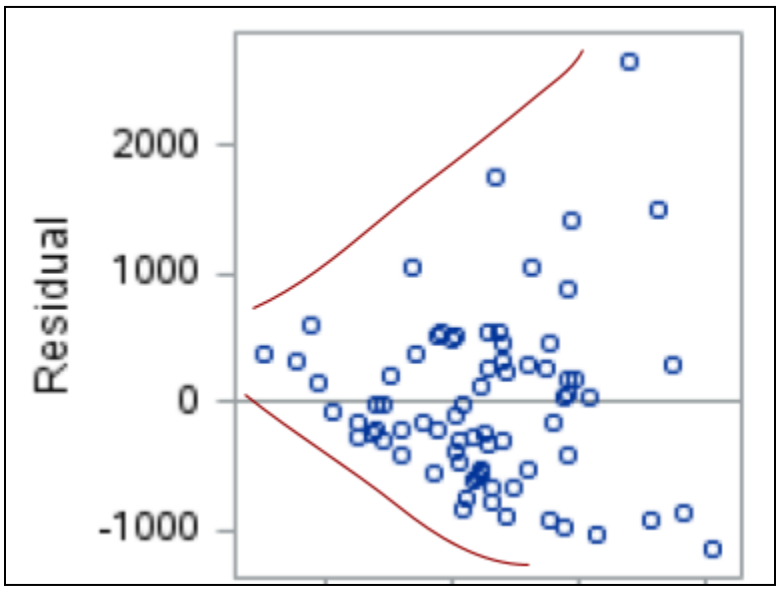
Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	**W**	0.934464	**Pr < W**	0.0008
Kolmogorov-Smirnov	**D**	0.095733	**Pr > D**	0.0914
Cramer-von Mises	**W-Sq**	0.113147	**Pr > W-Sq**	0.0775
Anderson-Darling	**A-Sq**	0.871913	**Pr > A-Sq**	0.0242

To double check the normality, we look at the shapiro-wilk p-value and do a hypothesis test.
Where,
H0: $\varepsilon i \sim$ Normal (The residuals follow a normal distribution)
H1: $\varepsilon i \not\sim$ Normal (The residuals follow a normal distribution)

P-Value: 0.0008
Decision: Since P-Value=0.0008 < Alpha = 0.05, we reject the null hypothesis
Conclusion: The residuals are not normally distributed

Check for Equal Variance

To check for equal variance, I observe the residual versus predicted value plot.

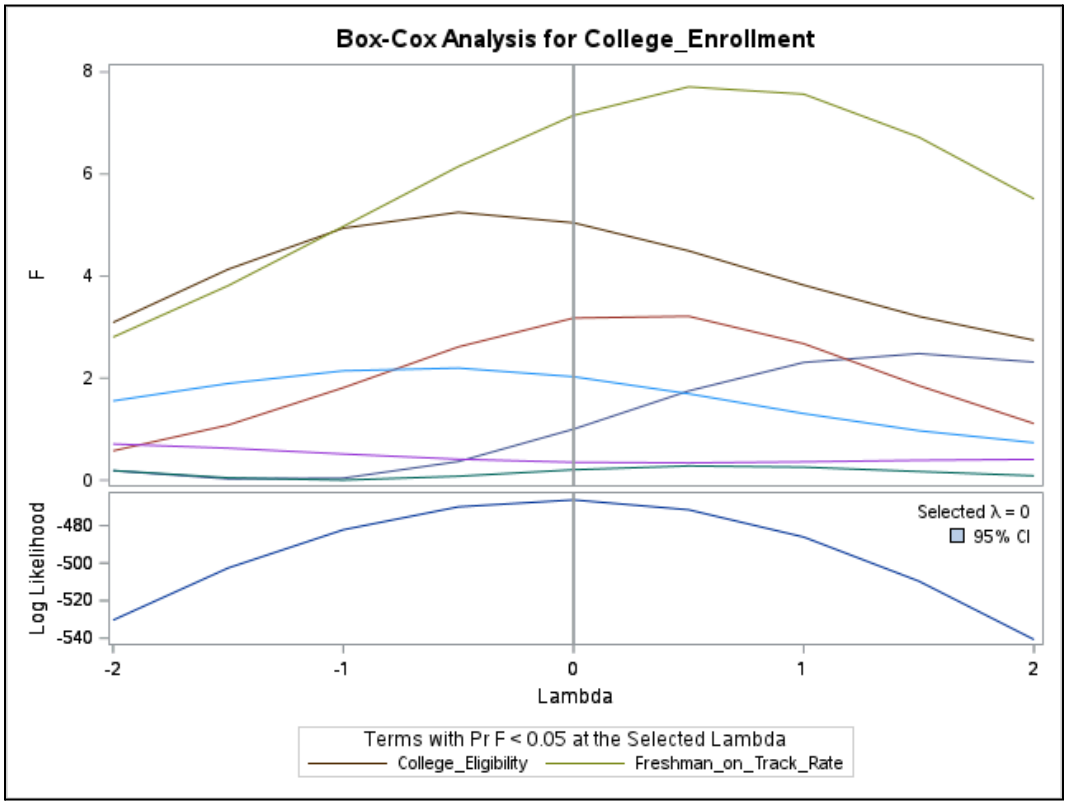


Based on the plot, there is a fanned pattern indicating that there is not equal variance.

Conclusion:

The model failed in the equal variance, normality, and linearity check. Since the equal variance and normality check failed, I will perform a Box-Cox transformation to change the college_enrollment (y) variable and see if it fixes our failed checks.

Model Transformation



The Box-Cox transformation procedure determines the best lambda value for the transformation.

Boxcox Transformation

The TRANSREG Procedure

TRANSREG Univariate Algorithm Iteration History for BoxCox(College_Enrollment)

Iteration Number	Average Change	Maximum Change	R-Square	Criterion Change	Note
1	0.00000	0.00000	0.21923		Converged

Algorithm converged.

Model Statement Specification Details

Type	DF	Variable	Description	Value
Dep	1	BoxCox(College_Enrollment)	Lambda Used	0
			Lambda	0
			Log Likelihood	-466.3
			Conv. Lambda	0
			Conv. Lambda LL	-466.3
			CI Limit	-468.2
			Alpha	0.05
			Options	Convenient Lambda Used
Ind	1	Identity(Average_Student_Attence)	DF	1
Ind	1	Identity(Rate_of_Misconducts)	DF	1
Ind	1	Identity(Average_Teacher_Attence)	DF	1
Ind	1	Identity(College_Eligibility)	DF	1
Ind	1	Identity(Graduation_Rate)	DF	1
Ind	1	Identity(Freshman_on_Track_Rate)	DF	1
Ind	1	Identity(Probation)	DF	1

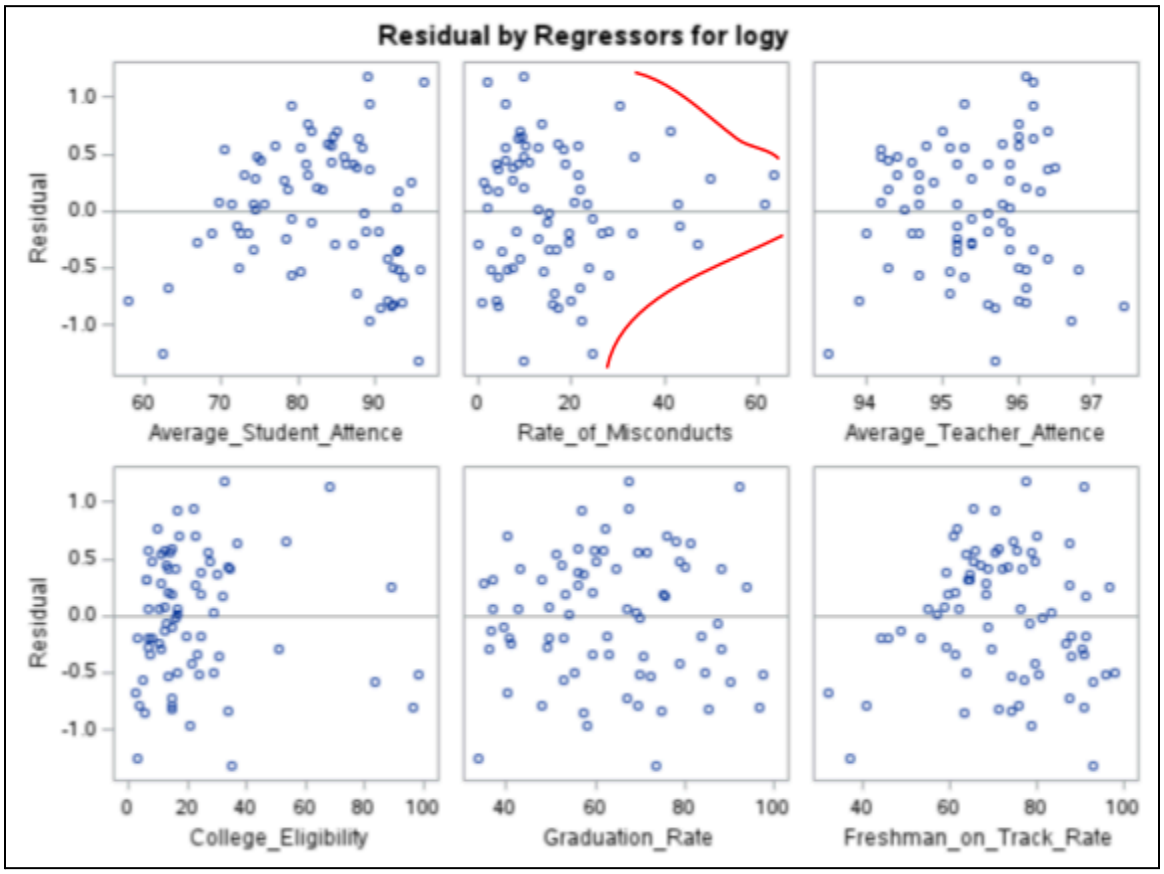
Based on the Box-Cox transformation plot, it seems that the selected lambda value is 0.

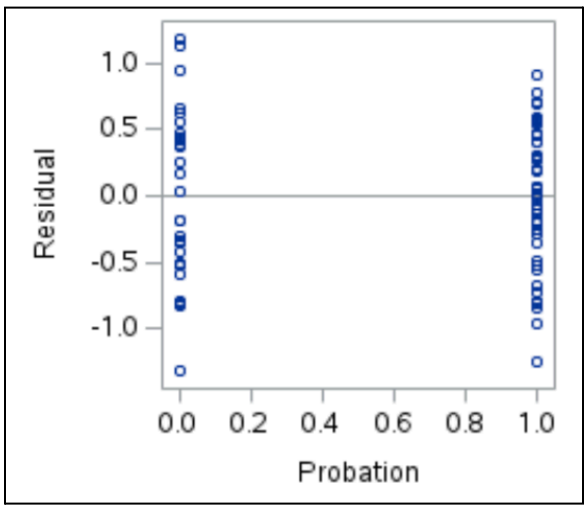
The provided charts confirm that the convenient lambda value for the transformation is 0.

Regression Model After Transformation

After the transformation, the model assumptions need to be rechecked.

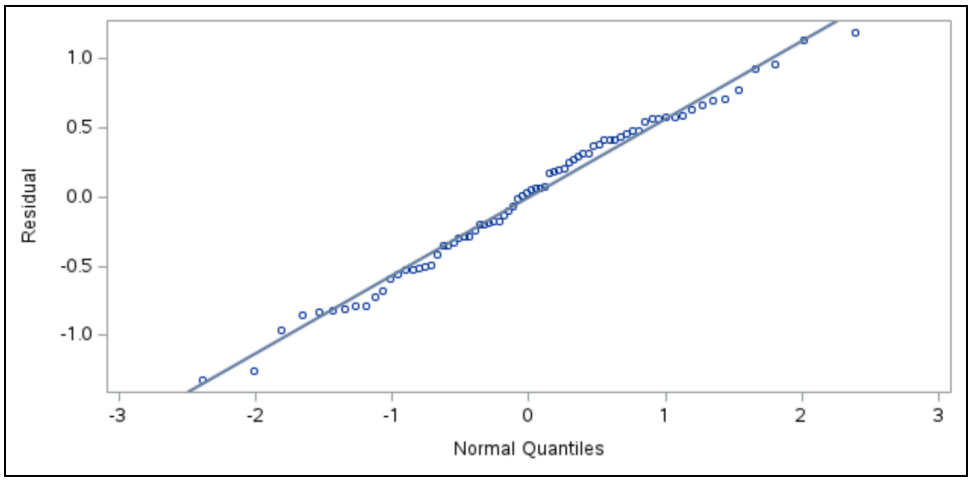
Check for Linearity





After the transformation, it appears that the X variable Rate_of_Misconduct still has a fanned pattern. Linearity check still fails.

Check for Normality



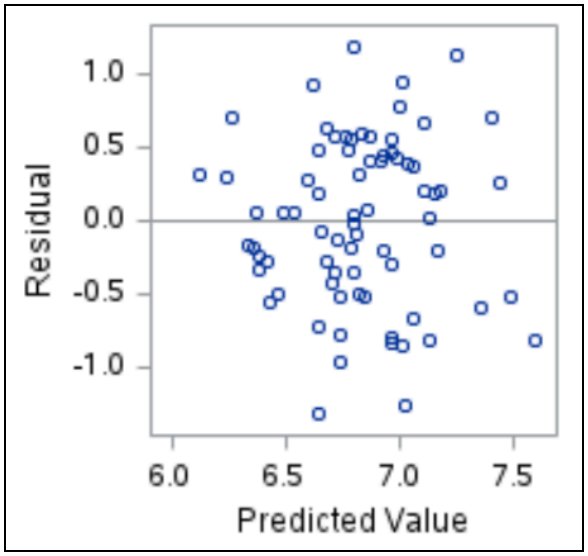
Checking the residuals versus normal quantiles plot, there does not seem to be any points that deviate. This indicates that the error of residuals are normal. To double check, we will do a shapiro-wilks test for normality.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.985045	Pr < W	0.5327
Kolmogorov-Smirnov	D	0.068332	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.062288	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.37657	Pr > A-Sq	>0.2500

Where,
H0: $\varepsilon i \sim$ Normal (The residuals follow a normal distribution)
H1: $\varepsilon i \not\sim$ Normal (The residuals follow a normal distribution)

P-Value: 0.5327
Decision: Since P-Value= 0.5327 > Alpha = 0.05, we do not reject the null hypothesis
Conclusion: The error of residuals are normally distributed

Check for Equal Variance



Based on the Residuals versus Predicted Value plot, there is no pattern present indicating that the equal variance check passed.

Conclusion:
The model passes the Normality and Equal Variance checks but still fails the Linearity check. This means that there should be a transformation performed on the X variable but for this project, we will keep going through with the analysis without another transformation.

Hypothesis Check on Overall Model

Let's check if the overall model is statistically significant.

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	7	6.53355	0.93336	2.65	0.0178
Error	66	23.26877	0.35256		
Corrected Total	73	29.80232			

Based on our Analysis of Variance (ANOVA) table:
Hypothesis:
H0: $\beta1=\beta2=\beta3=\beta6=\beta7=\beta8=\beta9=0$ (not significant)
H1: At least one βi != 0 (is significant)
Test Statistic: F = 2.65 & P-value = 0.0178
Decision: P-Value = 0.0178 < Alpha = 0.05, reject H0
Conclusion: The overall model is statistically significant.

Variable Selection

Now that the model is significant, we can test the predictive variables for their significance individually in order to build the "best" model possible. I will use backward selection with the criteria for inclusion being a significance of 0.10 or lower.
The column of interest is "Pr > F." Since our significance is 0.10 or lower, anything bigger than that needs to be eliminated. This process will be repeated one by one since the P-Values may change after one elimination. Once a variable has been deleted, it is gone forever. It cannot be added back in.

Step 0:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	11.21425	10.08547	0.43589	1.24	0.2702
Average_Student_Attence	0.01633	0.01633	0.35272	1.00	0.3209
Rate_of_Misconducts	-0.01197	0.00672	1.12033	3.18	0.0792
Average_Teacher_Attence	-0.04963	0.10975	0.07208	0.20	0.6526
College_Eligibility	0.01214	0.00540	1.77877	5.05	0.0280
Graduation_Rate	0.00437	0.00738	0.12365	0.35	0.5557
Freshman_on_Track_Rate	-0.02166	0.00810	2.52066	7.15	0.0094
Probation	0.33182	0.23300	0.71506	2.03	0.1591

Based on the chart, we see that Average_Teacher_Attence has the highest P-value over 0.10. It needs to be deleted.

Step 1:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.68084	1.08622	13.17866	37.83	<.0001
Average_Student_Attence	0.01337	0.01487	0.28176	0.81	0.3717
Rate_of_Misconducts	-0.01198	0.00668	1.12240	3.22	0.0772
College_Eligibility	0.01205	0.00537	1.75655	5.04	0.0280
Graduation_Rate	0.00493	0.00723	0.16167	0.46	0.4981
Freshman_on_Track_Rate	-0.02158	0.00805	2.50395	7.19	0.0092
Probation	0.33836	0.23116	0.74641	2.14	0.1479

Based on the chart, we see that Graduation_Rate has the highest P-value over 0.10. It needs to be deleted.

Step 2:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.80016	1.06777	14.01812	40.56	<.0001
Average_Student_Attence	0.01487	0.01465	0.35593	1.03	0.3138
Rate_of_Misconducts	-0.01343	0.00631	1.56658	4.53	0.0369
College_Eligibility	0.01351	0.00490	2.62434	7.59	0.0075
Freshman_on_Track_Rate	-0.02067	0.00791	2.36154	6.83	0.0110
Probation	0.32672	0.22962	0.69976	2.02	0.1593

Based on the chart, we see that Average_Student_Attence has the highest P-value over 0.10. It needs to be deleted.

Step 3:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7.75216	0.51006	79.87260	231.00	<.0001
Rate_of_Misconducts	-0.01444	0.00623	1.86050	5.38	0.0233
College_Eligibility	0.01361	0.00490	2.66494	7.71	0.0071
Freshman_on_Track_Rate	-0.01584	0.00631	2.17476	6.29	0.0145
Probation	0.23530	0.21125	0.42897	1.24	0.2692

Based on the chart, we see that Probation has the highest P-value over 0.10. It needs to be deleted.

Step 4:

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	8.03312	0.44408	113.53366	327.22	<.0001
Rate_of_Misconducts	-0.01211	0.00587	1.47445	4.25	0.0430
College_Eligibility	0.01139	0.00449	2.23606	6.44	0.0134
Freshman_on_Track_Rate	-0.01756	0.00613	2.84457	8.20	0.0055

All the variables left in the model are now significant at the 0.10 level.

The final statistical model is:

$$Y = \beta 0 + \beta 2X2 + \beta 6X6 + \beta 7X7 + \beta 8X8 + \beta 9X9 + \epsilon$$

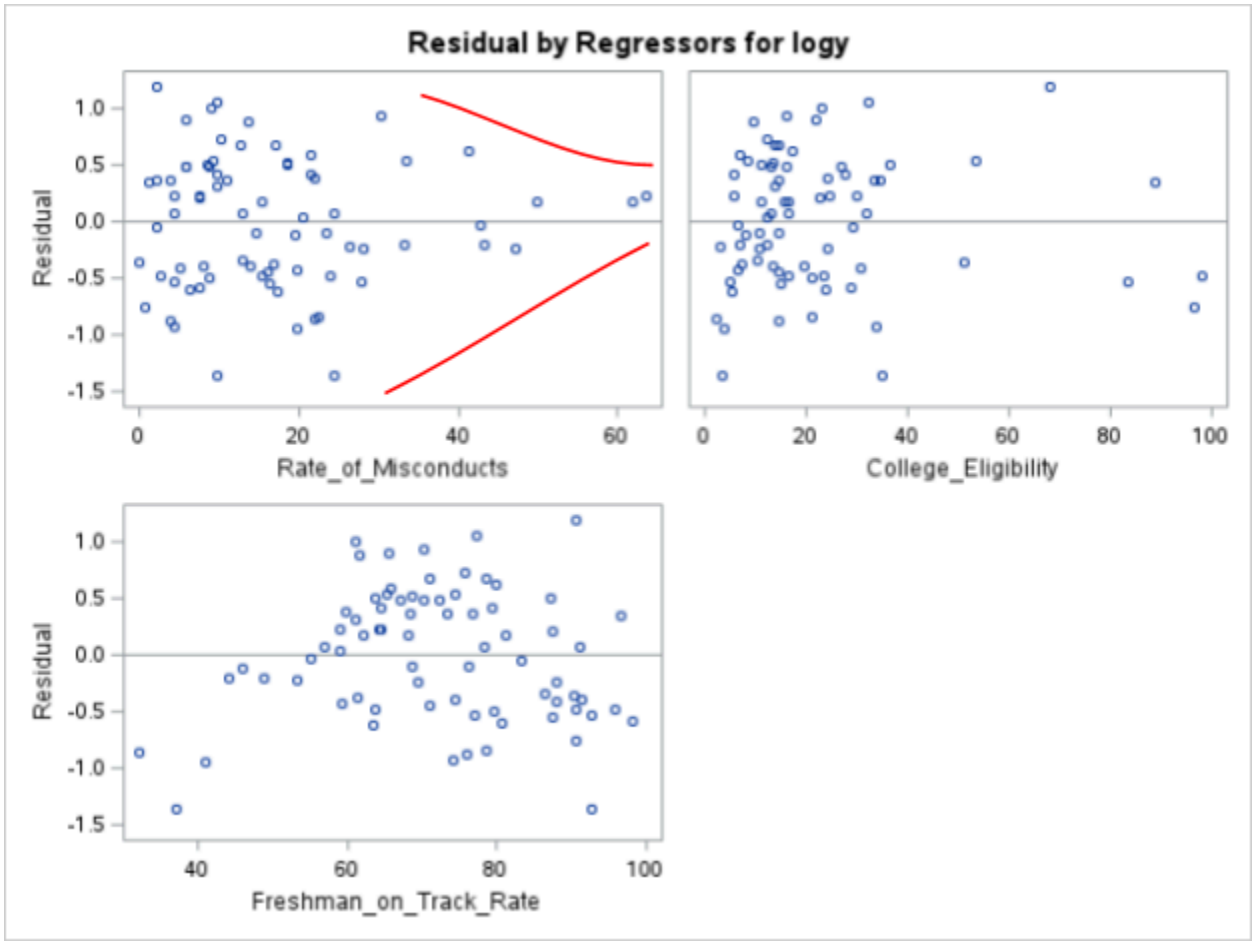
Or

$$Y = 8.033 - 0.012(Rate\_of\_Misconducts) + 0.011(College\_Eligibility) -$$

$$0.018(Freshman\_on\_Track\_Rate) + \epsilon$$

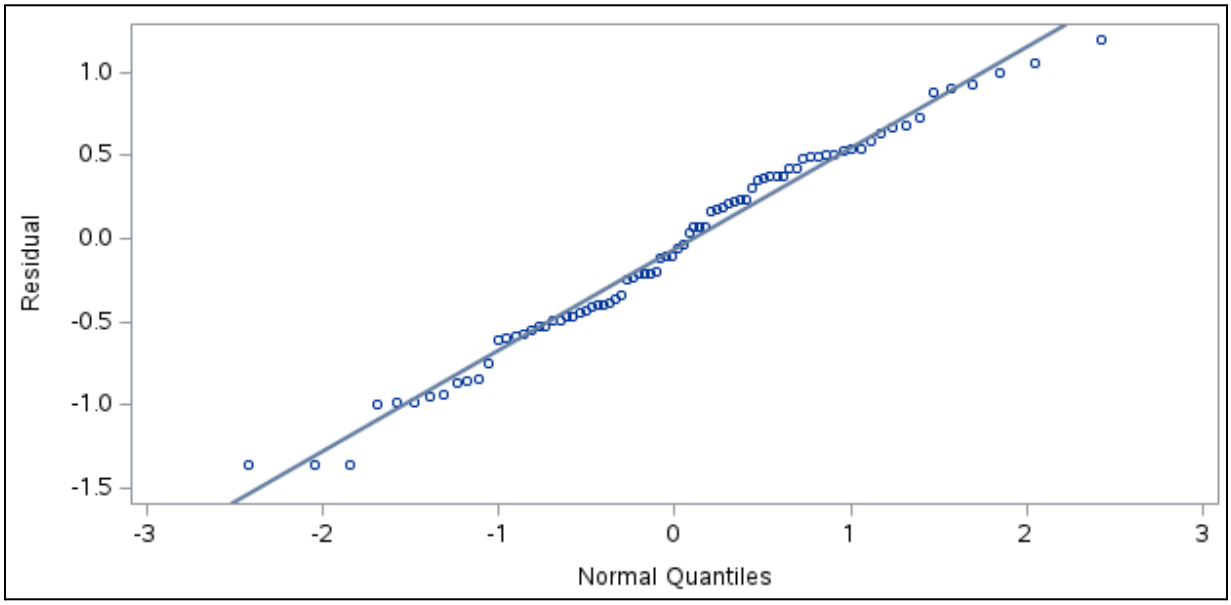
Final Model Assumptions Checks

Check for Linearity



Based on the Residual versus X*i* graphs, it appears that the predictor variable Rate_of_Misconducts still has its fanned pattern. As previously mentioned, a transformation of the X variable would be required.

Check for Normality



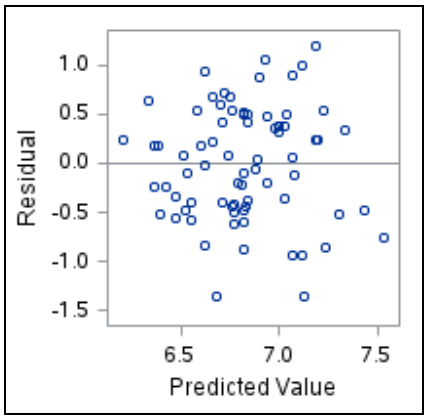
Based on the Residual versus Normal Quantiles plot, the points do not deviate from the line indicating that the error of residuals follow a normal distribution. We can follow up with a shapiro-wilk test.

Tests for Normality				
Test	**Statistic**		**p Value**	
Shapiro-Wilk	W	0.980983	Pr < W	0.2797
Kolmogorov-Smirnov	D	0.07525	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.081365	Pr > W-Sq	0.2037
Anderson-Darling	A-Sq	0.46604	Pr > A-Sq	0.2494

Where,
H0: $\varepsilon i \sim$ Normal (The residuals follow a normal distribution)
H1: $\varepsilon i \not\sim$ Normal (The residuals follow a normal distribution)

P-Value: 0.2797
Decision: Since P-Value= 0.2797 > Alpha = 0.05, we do not reject the null hypothesis
Conclusion: The error of residuals are normally distributed

Check for Equal Variance



The Residuals versus Predicted Value plot appears to have no pattern which indicates that the error of residuals have equal variance.

Conclusions

The final statistical model:
$Y = 8.033 - 0.012(Rate\_of\_Misconducts) + 0.011(College\_Eligibility) -$

$0.018(Freshman\_on\_Track\_Rate) + \epsilon$

Interpretations:

Rate of Misconducts: For each additional unit increase in the rate of misconducts, college enrollment is expected to decrease by 0.012 units, holding other factors constant.
College Eligibility: For each additional unit increase in college eligibility, college enrollment is expected to increase by 0.011 units.
Freshman on Track Rate: Each additional unit increase in freshman on track rate is associated with a decrease in college enrollment by 0.018 units.

Appendix

SAS code used below.

```
PROC IMPORT DATAFILE='/home/u63844046/Stats481/CPS_ES-HS_ProgressReport_2011-2012.csv'
    OUT=CPS_DATA
    DBMS=CSV
    REPLACE;
RUN;

PROC CONTENTS DATA=CPS_DATA; RUN;

/*----------------------------------Part I: Data Intro ----------------------------------*/
/* Descriptive Statistics for Continuous Variables */
PROC UNIVARIATE data = CPS_DATA;
    VAR Average_Student_Attence Rate_of_Misconducts
    Average_Teacher_Attence X9_Grade_Explore_2009
    X11_Grade_Average_ACT_2011 College_Eligibility
    Graduation_Rate Freshman_on_Track_Rate
    College_Enrollment;
    HISTOGRAM / NORMAL;
RUN;

/* Boxplots for Continuous Variables */
PROC SGPLOT DATA=CPS_DATA;
    VBOX Average_Student_Attence;
    VBOX Rate_of_Misconducts;
    VBOX Average_Teacher_Attence;
    VBOX X9_Grade_Explore_2009;
    VBOX X11_Grade_Average_ACT_2011;
    VBOX College_Eligibility;
    VBOX Graduation_Rate;
    VBOX Freshman_on_Track_Rate;
    VBOX College_Enrollment;
RUN;

/* Frequency Table for Indicator Variable */
PROC FREQ DATA=CPS_DATA;
    TABLES Probation;
RUN;

/* Missing Values in our Predictor Variables */
PROC MEANS DATA=CPS_DATA N NMISS;
    VAR Average_Student_Attence Rate_of_Misconducts Average_Teacher_Attence X9_Grade_Explore_2009
    X11_Grade_Average_ACT_2011 College_Eligibility Graduation_Rate Freshman_on_Track_Rate College_Enrollment
    Probation;
RUN;

/*------------------------------Part II: Multiple Linear Regression Analysis------------------------------*/

/* Multiple Linear Regression with VIF */
PROC REG DATA=CPS_DATA;
    MODEL College_Enrollment= Average_Student_Attence Rate_of_Misconducts
    Average_Teacher_Attence X9_Grade_Explore_2009
    X11_Grade_Average_ACT_2011 College_Eligibility
    Graduation_Rate Freshman_on_Track_Rate Probation / VIF;

RUN;

/* Removing X11_Grade_Average_ACT_2011 Due to High VIF - Rerunning Model*/
PROC REG DATA=CPS_DATA;
    MODEL College_Enrollment= Average_Student_Attence Rate_of_Misconducts
                              Average_Teacher_Attence X9_Grade_Explore_2009
                              College_Eligibility Graduation_Rate
                              Freshman_on_Track_Rate Probation / VIF;
RUN;

/* Removing X9_Grade_Explore_2009 Due to High VIF - Rerunning Model*/
PROC REG DATA=CPS_DATA;
    MODEL College_Enrollment= Average_Student_Attence Rate_of_Misconducts
                              Average_Teacher_Attence College_Eligibility
                              Graduation_Rate Freshman_on_Track_Rate Probation / VIF;
RUN;

/* ASSUMPTIONS CHECK */
PROC REG DATA=CPS_DATA;
    MODEL College_Enrollment= Average_Student_Attence Rate_of_Misconducts
                              Average_Teacher_Attence College_Eligibility
                              Graduation_Rate Freshman_on_Track_Rate Probation;
    OUTPUT OUT=result1 RESIDUAL=residual;
    TITLE 'Model';
RUN;
```

```
/* Normality of Residuals Check */
PROC UNIVARIATE DATA=result1 NORMAL PLOT;
    VAR residual;
RUN;

/* There is a violation with equal variance, normality of residuals, and linearity of x */
/* Do a Box-Cox Y transformation */
PROC TRANSREG DATA=CPS_DATA DETAIL;
MODEL BOXCOX(College_Enrollment / convenient lambda = -2 to 2 by 0.5)
    = identity(Average_Student_Attence Rate_of_Misconducts
               Average_Teacher_Attence College_Eligibility
               Graduation_Rate Freshman_on_Track_Rate Probation);
TITLE 'Boxcox Transformation';
RUN;

/* Perform a transformation on Y */
DATA CPS_DATA;
SET CPS_DATA;
logy = log(College_Enrollment);
RUN;

/* Full Model with All Variables After Transformation */
PROC REG DATA=CPS_DATA;
MODEL logy = Average_Student_Attence Rate_of_Misconducts
             Average_Teacher_Attence College_Eligibility
             Graduation_Rate Freshman_on_Track_Rate Probation;
OUTPUT OUT = result2 residual = residual;
TITLE 'Full Model After Transformation';
RUN;

PROC UNIVARIATE DATA=result2 NORMAL PLOT;
VAR residual;
RUN;

/* Hypothesis Test for Overall Model */
PROC REG DATA=CPS_DATA;
MODEL logy = Average_Student_Attence Rate_of_Misconducts
             Average_Teacher_Attence College_Eligibility
             Graduation_Rate Freshman_on_Track_Rate Probation;
TITLE 'Final Model';
RUN;

/*----------------------------------Part III: Variable Selection----------------------------------*/
/* Use Backward Elimination */
PROC REG DATA=CPS_DATA;
MODEL logy = Average_Student_Attence Rate_of_Misconducts
             Average_Teacher_Attence College_Eligibility
             Graduation_Rate Freshman_on_Track_Rate Probation
    / selection = backward CP SLSTAY = 0.10;
TITLE "BACKWARD SELECTION";
OUTPUT OUT = result3 residual = residual;
RUN;

PROC UNIVARIATE NORMAL PLOT DATA = result3;
    VAR residual;
    RUN;
```