2
(3 sections)

Analysis of Drop-Out Rate

Summary:

This aim of this analysis was to investigate factors directly controllable by a school board, and discover which, if any, influence student drop-out rate. To find such associations, multiple linear regression with backward elimination was used to analyze a pre-existing data set. Further, the comparative strengths of associated factors were computed using beta coefficients. The final results of the analysis showed that student enrollment and average teacher salary influence student drop-out rate.

Introduction:

Though academic failure is clearly a function of both an institute and its students, it seems reasonable that there exist factors directly controllable by an institute that can ameliorate a student's chance of academic success. This aim of this analysis is to learn whether such influential factors have been measured in a pre-existing dataset, measure the degree of their influence, and discuss future studies that could aid in predicting and affecting student drop-out rate.

The data set analyzed contained information on 135 public high schools in a large Northeastern US metropolitan area. The focus of this analysis is on student drop-out rate, as well as those factors directly controllable by schools which could influence the rate. Specifically, the variables from the dataset falling under this criterion included: Student Enrollment, the Cost Per Pupil, Average Teacher Salary, Student to Teacher Ratio, and Student to Counselor Ratio. Remaining variables were precluded on the basis of the question of interested. For instance, though data on student performance on standardized tests is available, it is not directly controllable by the school board. Analysis took the form of multiple linear regression, with explanatory variables being selected via backward elimination, and strength of influence being measured via beta coefficients.

Summary statistics and plots of variables of interest are provided below.

Variable	N	Mean	Std Dev	Minimum	Maximum
Enrollment	135	1103.39	572.4392425	266.0000000	3945.00
Cost_Per_Pupil	135	7111.96	1192.01	4675.00	12586.00
Average_Teacher_Salary	135	46963.23	4468.55	32067.00	66654.00
Student_Teacher_Ratio	135	13.1333333	1.9386101	9.0000000	19.0000000
Student_Counselor_Ratio	135	224.5185185	52.9978670	113.0000000	389.0000000
Dropout_Rate	135	2.0481481	2.2102830	0	12.2000000

Figure 1: Summary statistics for the variables of interest.

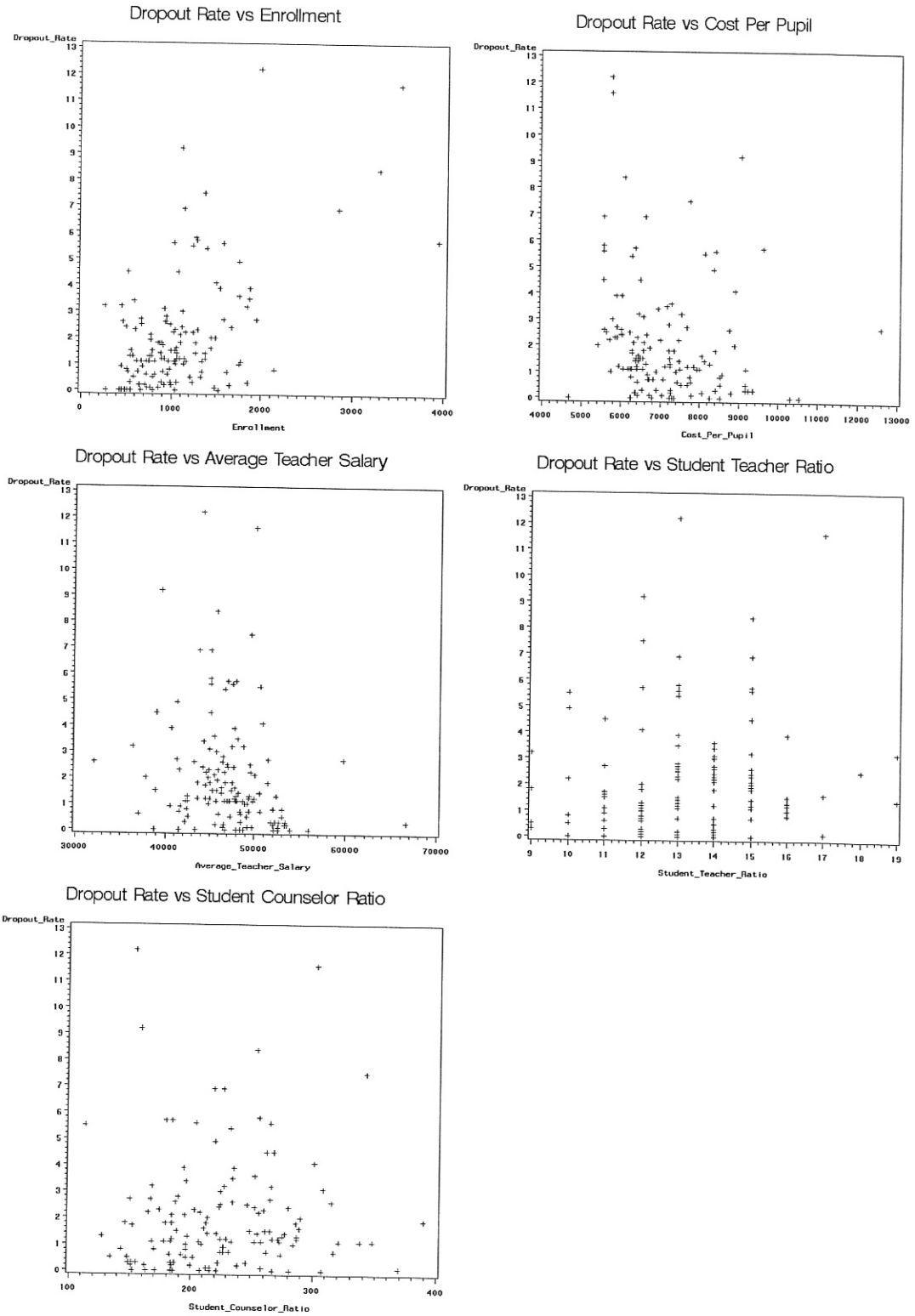

Figure 2: Scatter plots of Independent variables against Drop-Out Rate

Statistical Analysis:

The goal of this analysis was to discover those controllable factors that influenced student drop-out rate. In order to discover such influential factors, multiple linear regression was used with backwards elimination. Beta coefficients were then computed to facilitate comparing the strengths of these factors. This analysis approach was used as it is both common, and lends to results that are easy to understand.

To begin the analysis, we used the full regression model, which regressed student Drop-Out Rate against Enrollment, Cost Per Pupil, Average Teacher Salary, Student Teacher Ratio, and Student Counselor Ratio. However, not all explanatory variables in this model contributed significantly towards predicting drop-out rate. To address this issue, we employed the iterative backward elimination procedure. In this procedure, explanatory variables are dropped from the full model one at a time until all remaining explanatory variables make significant partial contributions towards predicting the dependent variables (Drop-Out Rate in our analysis). The variable deleted from the model at each iteration is the one that results in the smallest decrease in the Adjusted-R^2 value. Equivalently, the variable to be deleted can be described as the variable explaining the least variation in the dependent variable, while controlling for the predictors in the model. Proceeding in this manner, three variables were dropped from the full model. Specifically, Student Counselor Ratio, Student Teacher Ratio, and Cost Per Pupil were sequentially removed. Enrollment and Average Teacher Salary were the only predictor variables that explained a significant partial amount of the variability in the Drop-Out Rate and thus remained in the model. This model had an Adjusted-R^2 value of .38, specifying that after taking into account that there were two predictors in the model, 38% of the variability in Drop-Out Rate can be explained by the model. The coefficient for Average Teacher Salary was -0.00015220, while the coefficient for Enrollment was 0.00231. This result yields the following interpretation: For every unit increase in Enrollment, while holding the value of Average Teacher Salary constant, Drop-Out Rate is predicted to increase by 0.00231 units. Similarly, for every unit increase in Average Teacher Salary, while holding the value of Enrollment constant, Drop-Out Rate is predicted to decrease by 0.00015220. The reader should note that the values of these coefficients are not insignificant, as differences between student enrollment at schools often exceeded one thousand, and differences in average teacher salary between schools often exceeded ten thousand. The p-values for these estimates were both below 0.0001, having the interpretation that the values of coefficient estimates are highly significant, and extremely unlikely to have come by random chance.

As a next step in the analysis, we moved towards model diagnostics to see whether this model upheld the assumptions of multiple linear regression. This is a necessary step before any

regression model can be accepted as valid. First, we checked the normality of the residuals. Though the estimates of the coefficients of the regression equation are valid without this assumption holding, the assumption must hold in order for the p-values to be valid. A plot of the residuals is presented below, and quickly shows that they have positive skew and do not lend themselves to coming from a normal distribution. A secondary concern shown by the residuals plot was the possible existence of influential values. A standardized version of the residuals (called the studentized residuals) showed that there did exist several residuals which were above 3 standard deviations- a quality often associated with outliers. Reinvestigating the data showed that these observations did not represent outliers and they were influential to the model (as measured by Cooks Distance and Leverage), and thus could not be ignored.

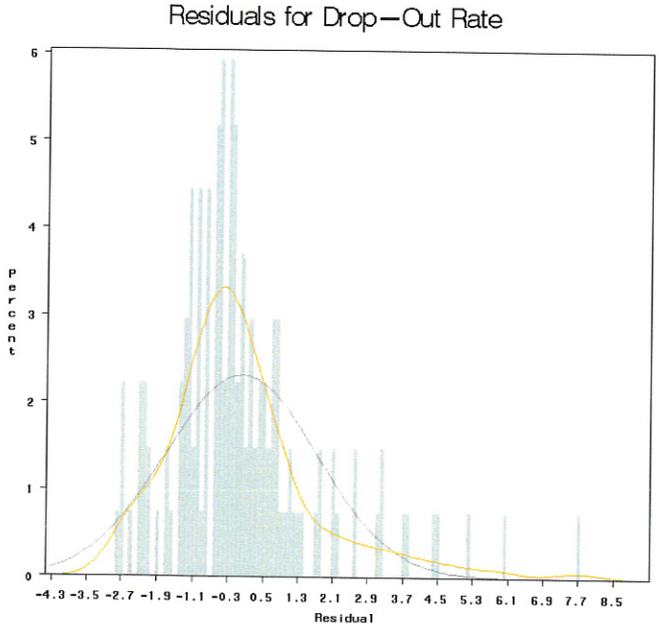

Figure 3: Residuals for Drop-Out Rate: Gray curve shows the normal curve that best fits the data. Orange curve shows kernel density which approximates the probability density of the residuals. Values of the residuals are shown in Green.

To overcome the issue of non-normal residuals, we next attempted to transform our depended variable Drop-Out Rate by taking its square root. Though changing the scale of the measurement for Drop-Out Rate does lead to a more complicated interpretation, it would allow for an analysis that did not violate the normal residuals assumption of our regression model, and is not incongruent with the goal of this analysis- to find factors that influenced Drop-Out Rate. Alternatives to this route do exist, but have their own trade-offs. For example, we could choose to take the log-transform of the dependent variable, to yield a loglinear model, or change our

regression model to one with a random component that is gamma distributed instead of normally distributed. However, both these approach add the complication of dealing with those observations for which Drop-Out Rate is zero (which is approximately 8% of the observations). As such, the approach of taking the square-root transformation was selected despite having the trade-off of having a more complicated interpretation.

A plot of the residuals of the Sqrt(Drop-Out Rate) is shown below, and shows a much more normal like distribution.

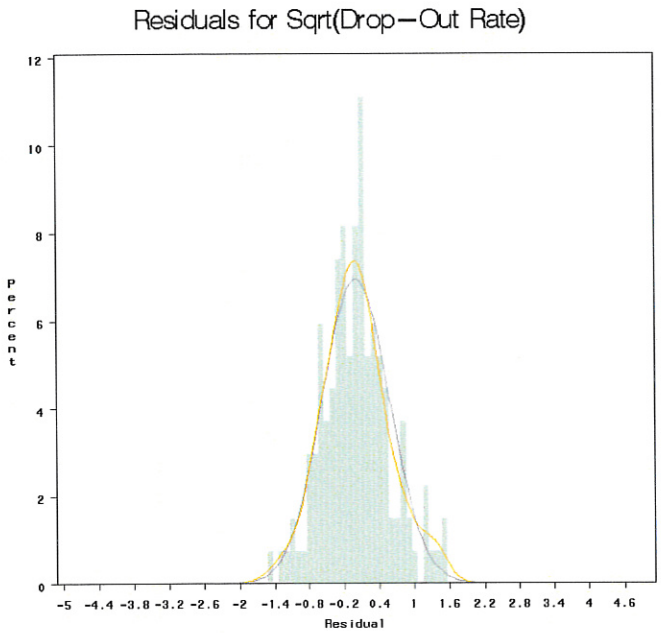

Figure 4: Residuals for Sqrt(Drop-Out Rate: Gray curve shows the normal curve that best fits the data. Orange curve shows kernel density which approximates the probability density of the residuals. Values of the residuals are shown in Green.

This square root transformed model had an Adjusted-R^2 value of .36, specifying that after taking into account that there were two predictors in the model, 36% of the variability in the Square-Root of the Drop-Out Rate can be explained by the model. The coefficient for Enrollment was 0.00071146, while the coefficient for Average Teacher Salary was -0.00005767. This result yields the following more complicated interpretation: For every unit increase in Enrollment, while holding the value of Average Teacher Salary constant, the Square-Root of the Drop-Out Rate is predicted to increase by 0.00071146 units. Similarly, for every unit increase in Average Teacher Salary, while holding the value of Enrollment constant, the Square-Root of the Drop-Out Rate is predicted to decrease by 0.00005767. Though the interpretation of these results is no longer straight forward, the results do show that Enrollment and Average Teacher Salary influence Drop-Out Rate (the p-values for both coefficients were below .0001).

The next assumption tested was that of homoscedasticity, specifically, that the standard deviation in our dependent variable is constant over all values of the explanatory variables. Our tool for this analysis is a plot of the residuals vs. predicted values for our dependent variable, and is shown below. Though extreme predicted values do seem to be closer together, this may simply be an artifact of the fact that there are so few values in the extremes (predicted value > 2 or predicted value < .5). As such, we find no reason to conclude that the homoscedasticity assumption is violated, but do note that this is a concern and could easily be addressed with more data.

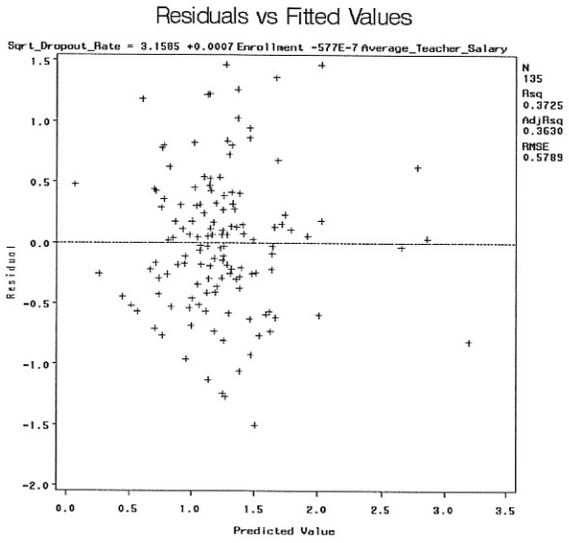

Figure 5: Plot of residuals vs. Predicted Values.

Finally, we searched for multicollinearlity between the Enrollment and Average Teacher Salary variables. The presence of multicollinearlity would indicate that the two predictor variables are nearly redundant, and thus one may not be needed in the model. However, regressing Enrollment against Average Teacher Salary yielded an R^2 value less that .03, indicting that multicollinearlity was not present.

Standardized beta coefficients were also computed for this regression model so that the relative strengths of the predictors could be compared. Specifically, the beta coefficient for Enrollment was .56145, and the beta coefficient for Average Teacher Salary was -.35524. This yields the following comparable interpretation: A one standard deviation increase in Enrollment, while holding the value of Average Teacher Salary constant, would yield a .56145 standard deviation increase in our dependent variable. Similarly, a one unit increase in Average Teacher Salary, while holding the value of Enrollment constant, would yield a .35524 standard deviation decrease in our dependant variable. The use of standardized beta coefficients allows us to

compare the relative strengths of the predictors in our regression model, and learn standardized deviations in Enrollment are more influential in predicting our dependant variable than equivalent standardized deviations in Average Teacher Salary.

Conclusions:

In this analysis we analyzed an existing dataset containing information on the drop out rate at different high schools along with several other measurements on those schools. From this dataset we were able to find two factors that influence drop out rate, student enrollment and average teacher salary.

These results, however, must be interpreted with care for several reasons. First, it is important to note that these associations do not imply causation. A school boards' attempt to lower enrollment and hire new teachers requesting higher salaries need not change drop out rate. Secondly, a larger dataset may show that the homoscedasticity assumption assumed in the model may not hold. As such, a new model would need to be built, and these data reanalyzed.

Finally we note that a large constraint is that the results are limited by the variables and observations we have data for. Future studies would benefit from choosing variables to answer the question of interest.

References:

- Casella & Berger, Statistical Inference, Second Edition, Chapters 11 & 12
- Agresti & Finlay, Statistical Methods for the Social Sciences, Third Edition, Chapter 14
- SAS 9.1

```
* EXPLORE THE DATA ================================================
* In this section we create a new data set called
"HS_DATA.Initial_Variables"
* containing only the variables of interest to our analysis. Then
we explore
* these variables by displaying their summary statistics and
plotting their
* histograms. This allows us to check for obvious errors in the
data;

* Import the data;
PROC IMPORT OUT= HS_DATA.All
   DATAFILE= "C:\Documents and Settings\aimee\Desktop\ksood
stats\highschools\schools.csv"
   DBMS=CSV REPLACE;
   GETNAMES=YES;
   DATAROW=2;

data HS_DATA.Initial_Variables;
   set HS_DATA.All;
   keep School Dropout_Rate Enrollment Cost_Per_Pupil
Average_Teacher_Salary
        Student_Teacher_Ratio Student_Counselor_Ratio;

PROC means data = HS_DATA.Initial_Variables;

proc univariate data = HS_DATA.Initial_Variables;
   var Dropout_Rate;
   histogram /cfill = gray midpoints=0 to 13 by .1;

proc univariate data = HS_DATA.Initial_Variables;
   var Enrollment;
   histogram /cfill = gray midpoints=0 to 4000 by 100;

proc univariate data = HS_DATA.Initial_Variables;
```

```
    var Cost_Per_Pupil;
    histogram /cfill = gray midpoints=4000 to 13000 by 100;

proc univariate data = HS_DATA.Initial_Variables;
    var Average_Teacher_Salary;
    histogram /cfill = gray midpoints=30000 to 70000 by 1000;

proc univariate data = HS_DATA.Initial_Variables;
    var Student_Teacher_Ratio;
    histogram /cfill = gray midpoints=5 to 20 by 1;

proc univariate data = HS_DATA.Initial_Variables;
    var Student_Counselor_Ratio;
    histogram /cfill = gray midpoints=100 to 400 by 5;

PROC GPLOT DATA=HS_DATA.Initial_Variables;
      PLOT Dropout_Rate * Enrollment;
      PLOT Dropout_Rate * Cost_Per_Pupil;
      PLOT Dropout_Rate * Average_Teacher_Salary;
      PLOT Dropout_Rate * Student_Teacher_Ratio;
      PLOT Dropout_Rate *Student_Counselor_Ratio;
RUN;
* ============================================================;
```

```
* MODEL DIAGNOSTICS (1) - EXAMINING THE RESIDUALS ==============
* In this section begin looking at the model prevously
constructed to see if
* any of the regression model assumptions are grossly violated.
To begin, we
* check that the residuals are normally distributed. Though the
estimates of
* the coefficients are valid without this assumption holding, the
assumption
* must hold in order for the results of the t-tests to be valid.;

* Create a new data set with this initial model and the
residuals;
PROC reg data = HS_DATA.Initial_Variables;
   model Dropout_Rate = Enrollment Average_Teacher_Salary;
   output out = HS_DATA.Model(keep = School
                              Dropout_Rate
                              Enrollment
                              Average_Teacher_Salary
                              Predicted_Dropout_Rate
                              Studentized_Residual
                              R
                              Leverage
                              Cooks_Distance
                              Dffits)
          P= Predicted_Dropout_Rate
          RSTUDENT= Studentized_Residual
          RESIDUAL = R
          H= Leverage
          COOKD= Cooks_Distance
          DFFITS= Dffits;

TITLE 'Residuals for Drop-Out Rate';
PROC univariate data = HS_DATA.Model;
   var R;
```

```
   histogram /cfill=CXBDCEBD cbarline=CXC6D9FD
normal(color=CX7E7E7E) kernel(color=CXFF9900)  midpoints= -4 to 8
by .1;

* The plot shows that there may be some influential observations,
and the
* resuiduals do not lend themselves to come from a normal
distribution.;
TITLE;
proc print data= HS_DATA.Model;
   where abs(Studentized_Residual) > 3;
   var School Studentized_Residual Cooks_Distance Leverage;
RUN;

* We further look at the residuals and some other statistics
* to see if there exist any outliers.;
proc print data= HS_DATA.Model;
   where Cooks_Distance > (4/135);
   var School Studentized_Residual Cooks_Distance Leverage;
RUN;

proc print data= HS_DATA.Model;
   where Leverage > (6/135);
   var School Studentized_Residual Cooks_Distance Leverage;
RUN;
* ============================================================;
```

```
* MODEL DIAGNOSTICS (3) - HOMOSCEDASTICITY & MULTICOLLINEARITY=;

* Plot the residual vs predicted values;
TITLE 'Homoscedasticity';
proc reg data = HS_DATA.Transformed_Data;
   model Sqrt_Dropout_Rate = Enrollment Average_Teacher_Salary;
   plot r.*p.;
RUN;

TITLE;
proc reg data = HS_DATA.Transformed_Data;
   model Average_Teacher_Salary = Enrollment;
RUN;
* ================================================================;
```

2

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Enrollment	135	1103.39	572.4392425	266.0000000	3945.00
Cost_Per_Pupil	135	7111.96	1192.01	4675.00	12586.00
Average_Teacher_Salary	135	46963.23	4468.55	32067.00	66654.00
Student_Teacher_Ratio	135	13.1333333	1.9386101	9.0000000	19.0000000
Student_Counselor_Ratio	135	224.5185185	52.9978670	113.0000000	389.0000000
Dropout_Rate	135	2.0481481	2.2102830	0	12.2000000

The UNIVARIATE Procedure

Variable: Dropout_Rate

Moments

N	135	Sum Weights	135
Mean	2.04814815	Sum Observations	276.5
Std Deviation	2.21028302	Variance	4.88535102
Skewness	2.16248584	Kurtosis	5.75420127
Uncorrected SS	1220.95	Corrected SS	654.637037
Coeff Variation	107.916169	Std Error Mean	0.19023087

Basic Statistical Measures

Location		Variability	
Mean	2.048148	Std Deviation	2.21028
Median	1.300000	Variance	4.88535
Mode	0.000000	Range	12.20000
		Interquartile Range	2.00000

NOTE: The mode displayed is the smallest of 2 modes with a count of 11.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	10.76664	Pr > |t|	<.0001
Sign	M	62	Pr >= |M|	<.0001
Signed Rank	S	3875	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	12.2
99%	11.6
95%	6.9
90%	5.4
75% Q3	2.6
50% Median	1.3
25% Q1	0.6
10%	0.1
5%	0.0
1%	0.0
0% Min	0.0

The UNIVARIATE Procedure
Variable: Dropout_Rate

Extreme Observations

----Lowest---- ----Highest---

Value	Obs	Value	Obs
0	132	7.5	101
0	128	8.4	81
0	125	9.2	26
0	99	11.6	57
0	91	12.2	53

The UNIVARIATE Procedure
Variable: Enrollment

Moments

N	135	Sum Weights	135
Mean	1103.38519	Sum Observations	148957
Std Deviation	572.439242	Variance	327686.686
Skewness	2.14711145	Kurtosis	7.17413342
Uncorrected SS	208266963	Corrected SS	43910016
Coeff Variation	51.8802727	Std Error Mean	49.2677256

Basic Statistical Measures

Location		Variability	
Mean	1103.385	Std Deviation	572.43924
Median	998.000	Variance	327687
Mode	778.000	Range	3679
		Interquartile Range	595.00000

NOTE: The mode displayed is the smallest of 7 modes with a count of 2.

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	22.3957	Pr > |t|	<.0001
Sign	M	67.5	Pr >= |M|	<.0001
Signed Rank	S	4590	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	3945
99%	3526
95%	1951
90%	1765
75% Q3	1333
50% Median	998
25% Q1	738
10%	548
5%	469
1%	281
0% Min	266

The UNIVARIATE Procedure
Variable: Enrollment

Extreme Observations

----Lowest----		----Highest---	

Value	Obs	Value	Obs
266	103	2144	83
281	76	2849	14
431	125	3295	81
453	50	3526	57
455	68	3945	17

The UNIVARIATE Procedure
Variable: Cost_Per_Pupil

Moments

N	135	Sum Weights	135
Mean	7111.96296	Sum Observations	960115
Std Deviation	1192.0085	Variance	1420884.26
Skewness	1.21635064	Kurtosis	2.68239726
Uncorrected SS	7018700811	Corrected SS	190398491
Coeff Variation	16.7606117	Std Error Mean	102.591757

Basic Statistical Measures

Location		Variability	
Mean	7111.963	Std Deviation	1192
Median	6778.000	Variance	1420884
Mode	5561.000	Range	7911
		Interquartile Range	1473

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	69.32295	Pr > |t|	<.0001
Sign	M	67.5	Pr >= |M|	<.0001
Signed Rank	S	4590	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	12586
99%	10537
95%	9167
90%	8757
75% Q3	7754
50% Median	6778
25% Q1	6281
10%	5846
5%	5584
1%	5417
0% Min	4675

The UNIVARIATE Procedure
Variable: Cost_Per_Pupil

Extreme Observations

----Lowest----		----Highest----	
Value	Obs	Value	Obs
4675	99	9345	19

5417	13	9601	17
5561	111	10306	32
5561	85	10537	128
5561	31	12586	22

The UNIVARIATE Procedure
Variable: Average_Teacher_Salary

Moments

N	135	Sum Weights	135
Mean	46963.2296	Sum Observations	6340036
Std Deviation	4468.54554	Variance	19967899.2
Skewness	0.28517609	Kurtosis	3.00117499
Uncorrected SS	3.00424E11	Corrected SS	2675698496
Coeff Variation	9.51498773	Std Error Mean	384.591166

Basic Statistical Measures

Location		Variability	
Mean	46963.23	Std Deviation	4469
Median	47015.00	Variance	19967899
Mode	45106.00	Range	34587
		Interquartile Range	4588

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	122.1121	Pr > |t|	<.0001
Sign	M	67.5	Pr >= |M|	<.0001
Signed Rank	S	4590	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	66654
99%	59751
95%	53277
90%	52244
75% Q3	49426
50% Median	47015
25% Q1	44838
10%	41549
5%	38983
1%	36410
0% Min	32067

The UNIVARIATE Procedure
Variable: Average_Teacher_Salary

Extreme Observations

-----Lowest----		----Highest----	
Value	Obs	Value	Obs
32067	9	53490	54
36410	68	53921	91
37065	42	55994	32
37835	13	59751	22
38809	62	66654	28

The UNIVARIATE Procedure
Variable: Student_Teacher_Ratio

Moments

N	135	Sum Weights	135
Mean	13.1333333	Sum Observations	1773
Std Deviation	1.93861006	Variance	3.75820896
Skewness	0.32671375	Kurtosis	0.47723715
Uncorrected SS	23789	Corrected SS	503.6
Coeff Variation	14.7609903	Std Error Mean	0.16684899

Basic Statistical Measures

Location		Variability	
Mean	13.13333	Std Deviation	1.93861
Median	13.00000	Variance	3.75821
Mode	13.00000	Range	10.00000
		Interquartile Range	2.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	78.71389	Pr > |t|	<.0001
Sign	M	67.5	Pr >= |M|	<.0001
Signed Rank	S	4590	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	19
99%	19
95%	16
90%	15
75% Q3	14
50% Median	13
25% Q1	12
10%	11
5%	10
1%	9
0% Min	9

The UNIVARIATE Procedure
Variable: Student_Teacher_Ratio

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
9	103	17	57
9	97	17	58
9	19	18	102
9	7	19	88
10	128	19	115

The UNIVARIATE Procedure
Variable: Student_Counselor_Ratio

Moments

N	135	Sum Weights	135
Mean	224.518519	Sum Observations	30310
Std Deviation	52.997867	Variance	2808.77391
Skewness	0.47172522	Kurtosis	0.04941471
Uncorrected SS	7181532	Corrected SS	376375.704
Coeff Variation	23.6051206	Std Error Mean	4.56133014

Basic Statistical Measures

Location		Variability	
Mean	224.5185	Std Deviation	52.99787
Median	222.0000	Variance	2809
Mode	184.0000	Range	276.00000
		Interquartile Range	78.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	49.22216	Pr > |t|	<.0001
Sign	M	67.5	Pr >= |M|	<.0001
Signed Rank	S	4590	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	389
99%	369
95%	317
90%	289
75% Q3	262
50% Median	222
25% Q1	184
10%	155
5%	149
1%	127
0% Min	113

The UNIVARIATE Procedure
Variable: Student_Counselor_Ratio

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
113	98	337	41
127	78	342	101
134	123	348	108
143	83	369	93
146	97	389	114

The REG Procedure
Model: MODEL1
Dependent Variable: Dropout_Rate

Number of Observations Read	135
Number of Observations Used	135

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.3914 and C(p) = 6.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	256.22355	51.24471	16.59	<.0001
Error	129	398.41349	3.08848		
Corrected Total	134	654.63704			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7.56712	2.36976	31.49173	10.20	0.0018
Enrollment	0.00233	0.00028071	212.77038	68.89	<.0001
Cost_Per_Pupil	-0.00010957	0.00016112	1.42816	0.46	0.4977
Average_Teacher_Salary	-0.00014287	0.00003729	45.32440	14.68	0.0002
Student_Teacher_Ratio	-0.04674	0.09435	0.75784	0.25	0.6212
Student_Counselor_Ratio	0.00005757	0.00314	0.00104	0.00	0.9854

Bounds on condition number: 1.6004, 32.891

Backward Elimination: Step 1

Variable Student_Counselor_Ratio Removed: R-Square = 0.3914 and C(p) = 4.0003

The REG Procedure
Model: MODEL1
Dependent Variable: Dropout_Rate

Backward Elimination: Step 1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	256.22251	64.05563	20.90	<.0001
Error	130	398.41453	3.06473		
Corrected Total	134	654.63704			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	7.58280	2.20162	36.35505	11.86	0.0008
Enrollment	0.00233	0.00027931	213.20084	69.57	<.0001
Cost_Per_Pupil	-0.00011029	0.00015555	1.54080	0.50	0.4796
Average_Teacher_Salary	-0.00014289	0.00003713	45.38740	14.81	0.0002
Student_Teacher_Ratio	-0.04645	0.09271	0.76952	0.25	0.6172

Bounds on condition number: 1.5032, 20.948

Backward Elimination: Step 2

Variable Student_Teacher_Ratio Removed: R-Square = 0.3902 and C(p) = 2.2495

Analysis of Variance

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

Model	3	255.45299	85.15100	27.94	<.0001
Error	131	399.18404	3.04721		
Corrected Total	134	654.63704			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.84655	1.63492	53.43859	17.54	<.0001
Enrollment	0.00229	0.00026911	221.34953	72.64	<.0001
Cost_Per_Pupil	-0.00007273	0.00013591	0.87270	0.29	0.5934
Average_Teacher_Salary	-0.00014505	0.00003678	47.40222	15.56	0.0001

The REG Procedure
Model: MODEL1
Dependent Variable: Dropout_Rate

Backward Elimination: Step 2

Bounds on condition number: 1.1876, 10.156

Backward Elimination: Step 3

Variable Cost_Per_Pupil Removed: R-Square = 0.3889 and C(p) = 0.5321

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	254.58030	127.29015	42.00	<.0001
Error	132	400.05674	3.03073		
Corrected Total	134	654.63704			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.64753	1.58775	53.12583	17.53	<.0001
Enrollment	0.00231	0.00026673	227.23106	74.98	<.0001
Average_Teacher_Salary	-0.00015220	0.00003417	60.13173	19.84	<.0001

Bounds on condition number: 1.0307, 4.123

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Student_Counselor_Ratio	4	0.0000	0.3914	4.0003	0.00	0.9854
2	Student_Teacher_Ratio	3	0.0012	0.3902	2.2495	0.25	0.6172
3	Cost_Per_Pupil	2	0.0013	0.3889	0.5321	0.29	0.5934

The REG Procedure
Model: MODEL1
Dependent Variable: Dropout_Rate

Number of Observations Read	135
Number of Observations Used	135

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	254.58030	127.29015	42.00	<.0001
Error	132	400.05674	3.03073		
Corrected Total	134	654.63704			

Root MSE	1.74090	R-Square	0.3889
Dependent Mean	2.04815	Adj R-Sq	0.3796
Coeff Var	84.99874		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	6.64753	1.58775	4.19	<.0001
Enrollment	1	0.00231	0.00026673	8.66	<.0001
Average_Teacher_Salary	1	-0.00015220	0.00003417	-4.45	<.0001

The UNIVARIATE Procedure
Variable: R (Residual)

Moments

N	135	Sum Weights	135
Mean	0	Sum Observations	0
Std Deviation	1.72785939	Variance	2.98549807
Skewness	1.53173269	Kurtosis	3.71523676
Uncorrected SS	400.056742	Corrected SS	400.056742
Coeff Variation	.	Std Error Mean	0.14871046

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	1.72786
Median	-0.27520	Variance	2.98550
Mode	.	Range	10.44540
		Interquartile Range	1.66758

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-15.5	Pr >= |M|	0.0096
Signed Rank	S	-800	Pr >= |S|	0.0789

Quantiles (Definition 5)

Quantile	Estimate
100% Max	7.680933
99%	5.989451
95%	3.681193
90%	2.149943
75% Q3	0.640514
50% Median	-0.275196
25% Q1	-1.027062
10%	-1.867942
5%	-2.261147

1%	-2.735863
0% Min	-2.764470

The UNIVARIATE Procedure
Variable: R (Residual)

Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
-2.76447	17	4.41950	57
-2.73586	99	4.49155	21
-2.73150	83	5.23855	101
-2.69187	19	5.98945	26
-2.49149	2	7.68093	53

The UNIVARIATE Procedure
Fitted Distribution for R

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	1.727859

Goodness-of-Fit Tests for Normal Distribution

Test		---Statistic----		-----p Value-----	
Kolmogorov-Smirnov	D	0.13207354	Pr > D	<0.010	
Cramer-von Mises	W-Sq	0.71223991	Pr > W-Sq	<0.005	
Anderson-Darling	A-Sq	4.00319263	Pr > A-Sq	<0.005	

Quantiles for Normal Distribution

	------Quantile------	
Percent	Observed	Estimated
1.0	-2.73586	-4.019602
5.0	-2.26115	-2.842076
10.0	-1.86794	-2.214341
25.0	-1.02706	-1.165423
50.0	-0.27520	-0.000000
75.0	0.64051	1.165423
90.0	2.14994	2.214341
95.0	3.68119	2.842076
99.0	5.98945	4.019602

Obs	School	Studentized_ Residual	Cooks_ Distance	Leverage
26	Chelsea	3.65172	0.12324	0.029424
53	Lawrence	4.85053	0.21733	0.031422
101	Revere	3.12449	0.03372	0.010929
17	Brockton	-1.78385	0.25191	0.19447
26	Chelsea	3.65172	0.12324	0.02942
28	Concard-Carlisle	0.84921	0.04513	0.15779
53	Lawrence	4.85053	0.21733	0.03142
57	Lowell	2.80984	0.41084	0.14109
83	Newton North	-1.61133	0.03597	0.04037
101	Revere	3.12449	0.03372	0.01093

9	Avon	-0.15029	0.00077	0.09252
13	Billerica Memorial	-1.32237	0.02773	0.04566
14	BMC Durfee(Fall River)	0.19876	0.00129	0.08873
17	Brockton	-1.78385	0.25191	0.19447
22	Cambridge Rindge and Latin	0.38151	0.00402	0.07601
28	Concard-Carlisle	0.84921	0.04513	0.15779
32	Dover-Sherborn Regional	0.48256	0.00452	0.05470
42	Hamilton-Wenham Regional	-1.23060	0.02347	0.04459
57	Lowell	2.80984	0.41084	0.14109
68	Maynard	0.61431	0.00708	0.05307
81	New Bedford	0.67784	0.02166	0.12345

The REG Procedure
Model: MODEL1
Dependent Variable: Log_Dropout_Rate

Number of Observations Read	135
Number of Observations Used	124
Number of Observations with Missing Values	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	38.59929	19.29965	24.95	<.0001
Error	121	93.61458	0.77367		
Corrected Total	123	132.21387			

Root MSE	0.87959	R-Square	0.2919
Dependent Mean	0.34747	Adj R-Sq	0.2802
Coeff Var	253.13771		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	4.02861	0.85038	4.74	<.0001
Enrollment	1	0.00080479	0.00014168	5.68	<.0001
Average_Teacher_Salary	1	-0.00009819	0.00001849	-5.31	<.0001

The UNIVARIATE Procedure
Variable: R (Residual)

Moments

N	124	Sum Weights	124
Mean	0	Sum Observations	0
Std Deviation	0.8724071	Variance	0.76109415
Skewness	-0.5531849	Kurtosis	0.35117346
Uncorrected SS	93.6145803	Corrected SS	93.6145803
Coeff Variation	.	Std Error Mean	0.07834447

Basic Statistical Measures

Location		Variability	
Mean	0.000000	Std Deviation	0.87241
Median	0.133472	Variance	0.76109
Mode	.	Range	4.44613
		Interquartile Range	1.03372

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	8	Pr >= |M|	0.1777
Signed Rank	S	227	Pr >= |S|	0.5734

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.752561
99%	1.640270
95%	1.261354
90%	1.122656
75% Q3	0.506180
50% Median	0.133472
25% Q1	-0.527536
10%	-1.111619
5%	-1.676149
1%	-2.356715
0% Min	-2.693569

The UNIVARIATE Procedure
Variable: R (Residual)

Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
-2.69357	2	1.41675	21
-2.35672	93	1.54103	96
-2.07229	64	1.58766	103
-1.86375	45	1.64027	98
-1.86241	37	1.75256	101

Missing Values

Missing Value	Count	-----Percent Of----- All Obs	Missing Obs
.	11	8.15	100.00

The UNIVARIATE Procedure
Fitted Distribution for R

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	0.872407

Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic----		-----p Value-----	
Kolmogorov-Smirnov	D	0.07763462	Pr > D	0.067
Cramer-von Mises	W-Sq	0.12075029	Pr > W-Sq	0.061
Anderson-Darling	A-Sq	0.73194180	Pr > A-Sq	0.056

Quantiles for Normal Distribution

Percent	------Quantile------ Observed	Estimated
1.0	-2.35672	-2.029522
5.0	-1.67615	-1.434982
10.0	-1.11162	-1.118035
25.0	-0.52754	-0.588430
50.0	0.13347	-0.000000
75.0	0.50618	0.588430
90.0	1.12266	1.118035
95.0	1.26135	1.434982
99.0	1.64027	2.029522

The REG Procedure
Model: MODEL1
Dependent Variable: Sqrt_Dropout_Rate

Number of Observations Read	135
Number of Observations Used	135

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	26.26674	13.13337	39.18	<.0001
Error	132	44.24295	0.33517		
Corrected Total	134	70.50969			

Root MSE	0.57894	R-Square	0.3725
Dependent Mean	1.23525	Adj R-Sq	0.3630
Coeff Var	46.86823		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.15846	0.52801	5.98	<.0001
Enrollment	1	0.00071146	0.00008870	8.02	<.0001
Average_Teacher_Salary	1	-0.00005767	0.00001136	-5.07	<.0001

The UNIVARIATE Procedure
Variable: R (Residual)

Moments

N	135	Sum Weights	135
Mean	0	Sum Observations	0
Std Deviation	0.57460532	Variance	0.33017128
Skewness	0.23246894	Kurtosis	0.25279254
Uncorrected SS	44.2429514	Corrected SS	44.2429514
Coeff Variation	.	Std Error Mean	0.04945415

Basic Statistical Measures

Location		Variability	
Mean	0.00000	Std Deviation	0.57461
Median	-0.02363	Variance	0.33017
Mode	.	Range	2.97170
		Interquartile Range	0.68214

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	0	Pr > |t|	1.0000
Sign	M	-0.5	Pr >= |M|	1.0000
Signed Rank	S	-131	Pr >= |S|	0.7748

Quantiles (Definition 5)

Quantile	Estimate
100% Max	1.4656691
99%	1.4651312
95%	1.1822246
90%	0.8005829
75% Q3	0.3191931
50% Median	-0.0236275
25% Q1	-0.3629465
10%	-0.7087586
5%	-0.9213094
1%	-1.2676605
0% Min	-1.5060329

The UNIVARIATE Procedure
Variable: R (Residual)

Extreme Observations

------Lowest-----		-----Highest-----	
Value	Obs	Value	Obs
-1.50603	99	1.22346	96
-1.26766	62	1.26051	21
-1.24567	16	1.35839	26
-1.13042	69	1.46513	53
-1.06164	2	1.46567	101

The UNIVARIATE Procedure
Fitted Distribution for R

Parameters for Normal Distribution

Parameter	Symbol	Estimate
Mean	Mu	0
Std Dev	Sigma	0.574605

Goodness-of-Fit Tests for Normal Distribution

Test	---Statistic----		-----p Value-----	
Kolmogorov-Smirnov	D	0.06469400	Pr > D	>0.150
Cramer-von Mises	W-Sq	0.07179989	Pr > W-Sq	>0.250
Anderson-Darling	A-Sq	0.46744564	Pr > A-Sq	0.250

Quantiles for Normal Distribution

	------Quantile------	
Percent	Observed	Estimated
1.0	-1.26766	-1.336732
5.0	-0.92131	-0.945142

10.0	-0.70876	-0.736386
25.0	-0.36295	-0.387565
50.0	-0.02363	-0.000000
75.0	0.31919	0.387565
90.0	0.80058	0.736386
95.0	1.18222	0.945142
99.0	1.46513	1.336732

The REG Procedure
Model: MODEL1
Dependent Variable: Sqrt_Dropout_Rate

Number of Observations Read	135
Number of Observations Used	135

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	26.26674	13.13337	39.18	<.0001
Error	132	44.24295	0.33517		
Corrected Total	134	70.50969			

Root MSE	0.57894	R-Square	0.3725
Dependent Mean	1.23525	Adj R-Sq	0.3630
Coeff Var	46.86823		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	3.15846	0.52801	5.98	<.0001	0
Enrollment	1	0.00071146	0.00008870	8.02	<.0001	0.56145
Average_Teacher_Salary	1	-0.00005767	0.00001136	-5.07	<.0001	-0.35524

The REG Procedure
Model: MODEL1
Dependent Variable: Sqrt_Dropout_Rate

Number of Observations Read	135
Number of Observations Used	135

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	26.26674	13.13337	39.18	<.0001
Error	132	44.24295	0.33517		
Corrected Total	134	70.50969			

Root MSE	0.57894	R-Square	0.3725
Dependent Mean	1.23525	Adj R-Sq	0.3630
Coeff Var	46.86823		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	3.15846	0.52801	5.98	<.0001

Enrollment	1	0.00071146	0.00008870	8.02	<.0001
Average_Teacher_Salary	1	-0.00005767	0.00001136	-5.07	<.0001

The REG Procedure
Model: MODEL1
Dependent Variable: Average_Teacher_Salary

Number of Observations Read	135
Number of Observations Used	135

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	79808819	79808819	4.09	0.0452
Error	133	2595889677	19517967		
Corrected Total	134	2675698496			

Root MSE	4417.91438	R-Square	0.0298
Dependent Mean	46963	Adj R-Sq	0.0225
Coeff Var	9.40718		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	45476	828.09241	54.92	<.0001
Enrollment	1	1.34817	0.66671	2.02	0.0452