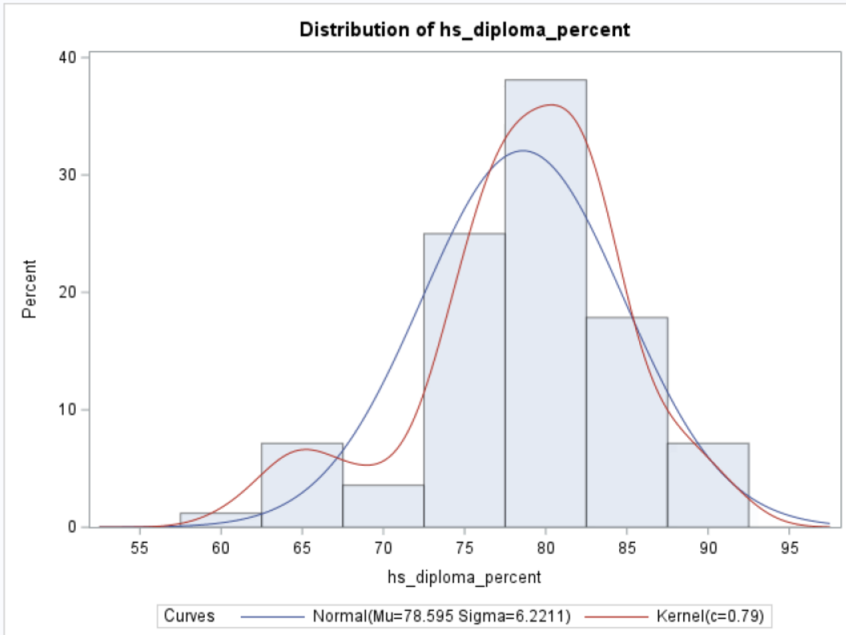
4. (SAS Exercise) Use the crime rate data described in KNNL Problem 1.28.

(a) Describe the distribution of the explanatory variable.



As evidenced by the histogram, the explanatory variable, the percentage of individuals in a selected county with at least a high school diploma, follows an approximately normal distribution with a slight left skew.

(b) Run the linear regression to predict the county crime rate from the percentage of individuals having at least a high school diploma.

The REG Procedure
Model: MODEL1
Dependent Variable: crime_rate

Number of Observations Read	84
Number of Observations Used	84

Analysis of Variance

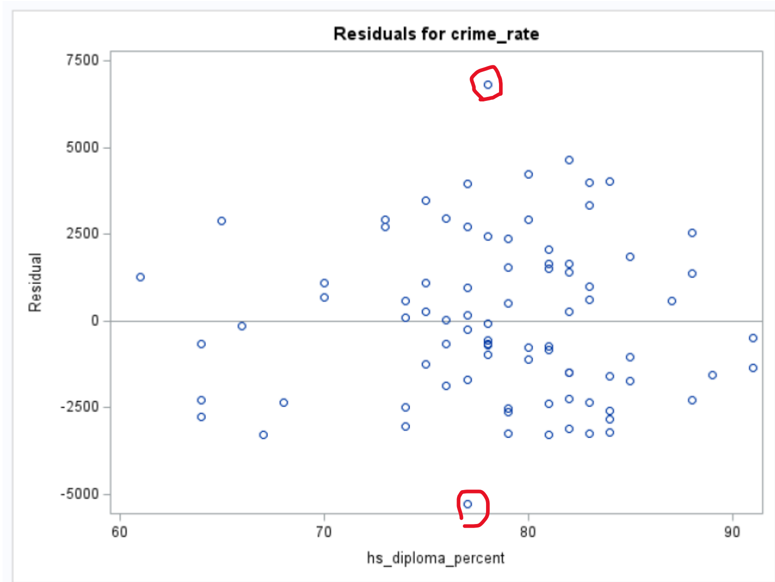
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	93462942	93462942	16.83	<.0001
Error	82	455273165	5552112		
Corrected Total	83	548736108			

Root MSE	2356.29195	R-Square	0.1703
Dependent Mean	7111.20238	Adj R-Sq	0.1602
Coeff Var	33.13493		

Parameter Estimates

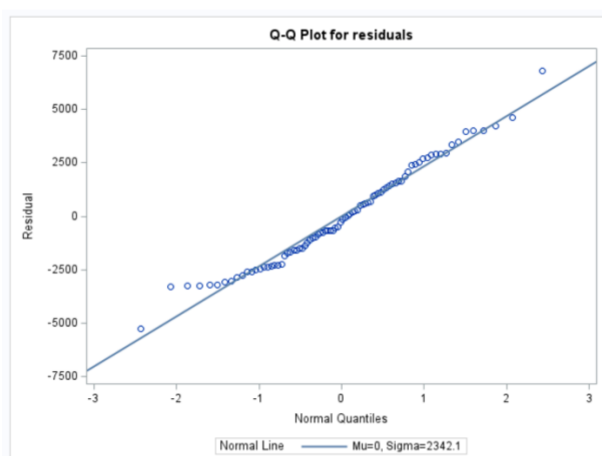
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20518	3277.64269	6.26	<.0001
hs_diploma_percent	1	-170.57519	41.57433	-4.10	<.0001

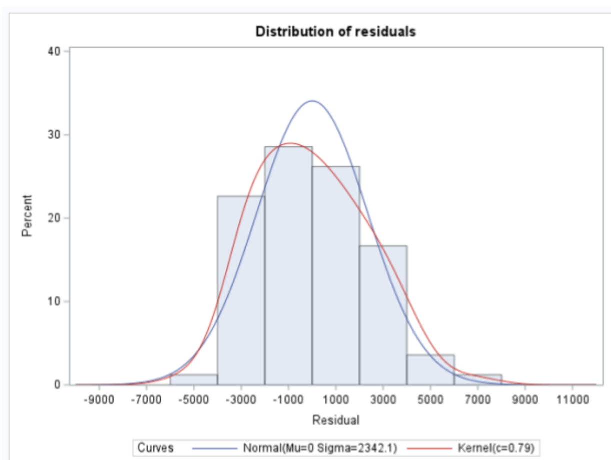
(c) Plot the residuals versus the explanatory variable and briefly describe the plot noting any unusual patterns or points.



The residuals appear to be randomly distributed. There are two outlier points circled in red corresponding to the maximum and minimum crime rate data points.

(d) Examine the distribution of the residuals by getting a histogram and a normal probability plot of the residuals by using the HISTOGRAM and QQPLOT statements in PROC UNIVARIATE. What do you conclude?





The distribution of the residuals is approximately normal. It is reasonable to conclude that the residuals are indeed normally distributed because they mostly follow a straight 45-degree line, as evidenced by the QQ plot.

5. (SAS Exercise) Use the crime rate data described in KNNL Problem 1.28. Change the data set by changing the value of the crime rate for the last observation from 7582 to 758 (e.g., a typo). You can do this in a data step.

(a) Make a table comparing the results of this analysis with the results of the analysis of the original data. Include in the table the following: fitted equation, t-test for the slope, with standard error and p-value, R2, and the estimate of σ2. Briefly summarize the differences.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	93462942	93462942	16.83	<.0001
Error	82	455273165	5552112		
Corrected Total	83	548736108			

Root MSE	2356.29195	R-Square	0.1703
Dependent Mean	7111.20238	Adj R-Sq	0.1602
Coeff Var	33.13493		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20518	3277.64269	6.26	<.0001
hs_diploma_percent	1	-170.57519	41.57433	-4.10	<.0001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	87518841	87518841	14.33	0.0003
Error	82	500804428	6107371		
Corrected Total	83	588323269			

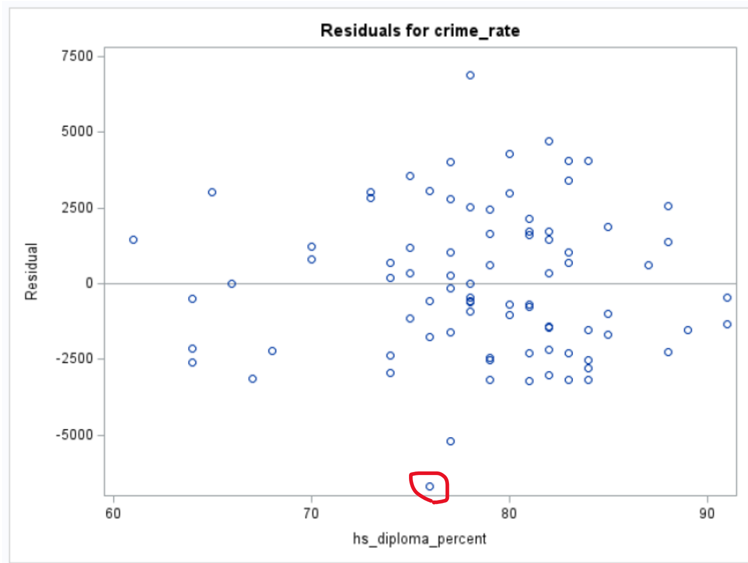
Root MSE	2471.30959	R-Square	0.1488
Dependent Mean	7029.96429	Adj R-Sq	0.1384
Coeff Var	35.15394		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20003	3437.63420	5.82	<.0001
hs_diploma_percent	1	-165.06193	43.60369	-3.79	0.0003

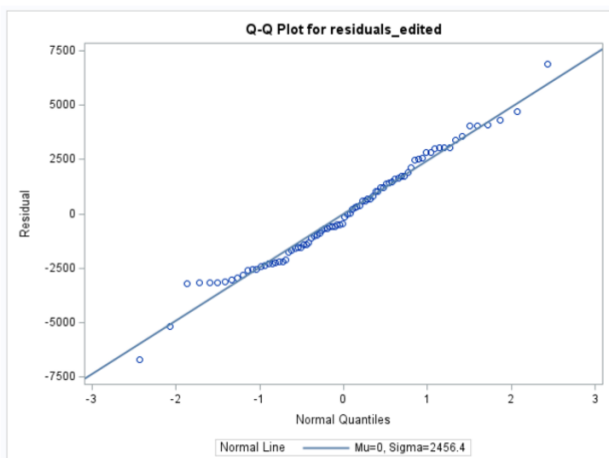
	Fitted Equation	t-test for slope	Standard Error	p-value	R^2	σ^2 Estimate
Original Data	$\hat{Y} = -170.575X_i + 20518$	-4.10	41.57433	<0.0001	0.1703	5552112
Corrected Data	$\hat{Y} = -165.017X_i + 20003$	-3.79	43.60369	0.0003	0.1488	6107371

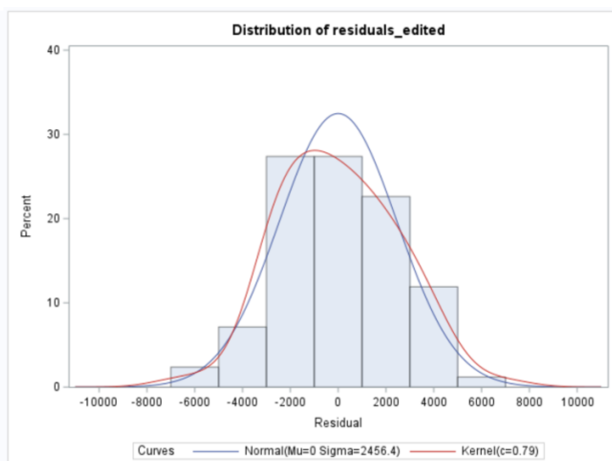
After changing the last observation (76, 7582) to (76, 758), the original data shows stronger significance in correlation when compared to the corrected data, as evidenced by the difference in t-tests, p-values, and R^2 values.

(b) Repeat parts (c) and (d) from the previous problem for this altered data set analysis and summarize how these plots help you to detect the unusual observation.



After examining the altered data's residual plot, there is now a new minimum crime rate highlighted in red. It corresponds to changing the last observation (76, 7582) to (76, 758).

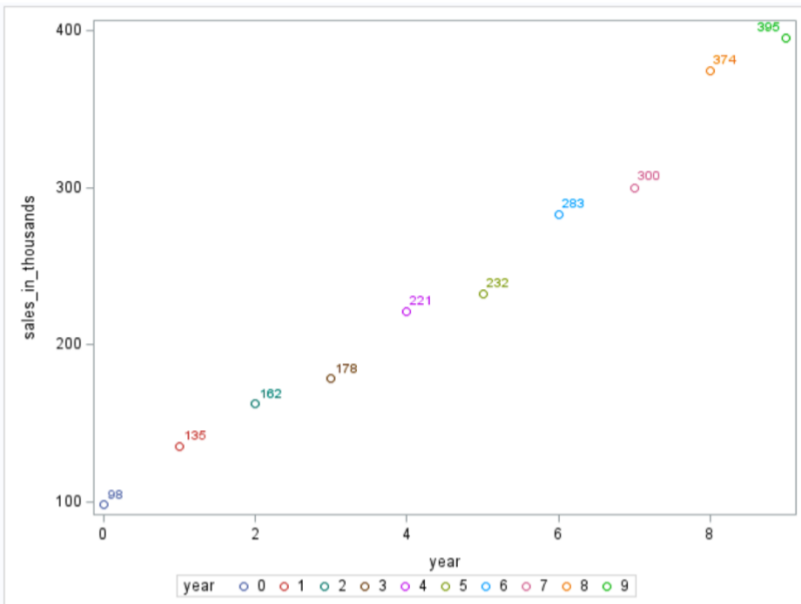




We see a stronger left skew in the data as evidenced by the histogram of residuals and the lower tail of the QQ plot.

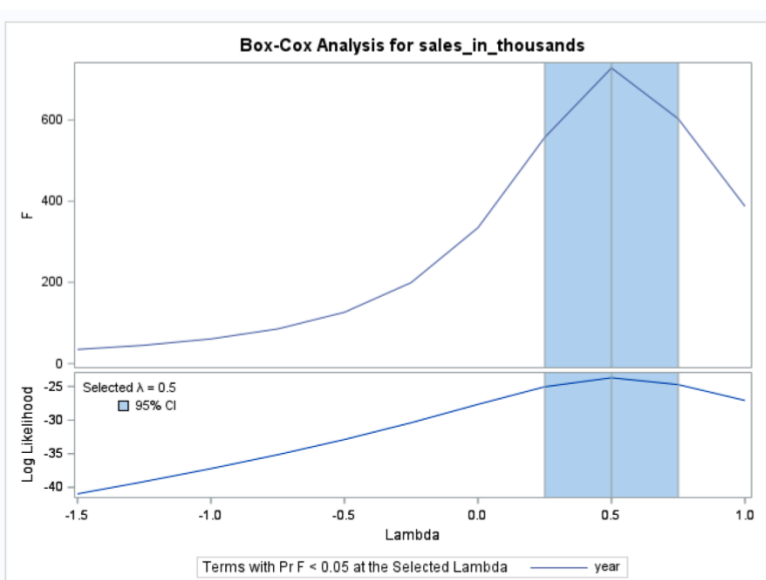
6. (SAS Exercise) Use the sales growth data described in KNNL Problem 3.17.

(a) Generate a scatterplot of the data and discuss the appropriateness of using a linear regression model.



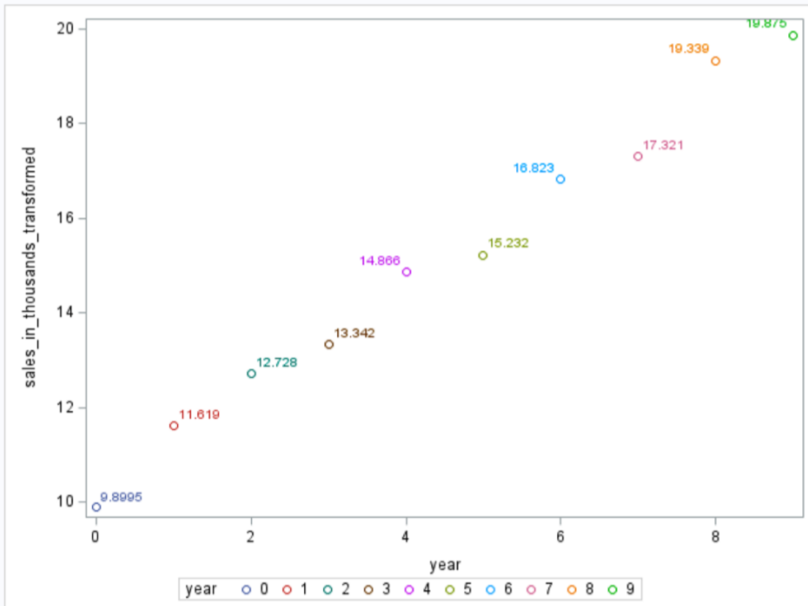
The data seems to follow a general linear trend, so a linear regression model is appropriate.

(b) Using PROC TRANSREG, which power transformation of Y (i.e., value of λ) is most appropriate to use here?



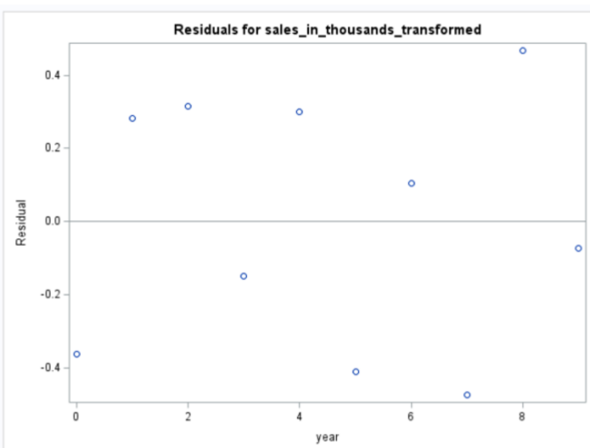
According to the Box-Cox plot, a λ value of 0.5 is the most appropriate.

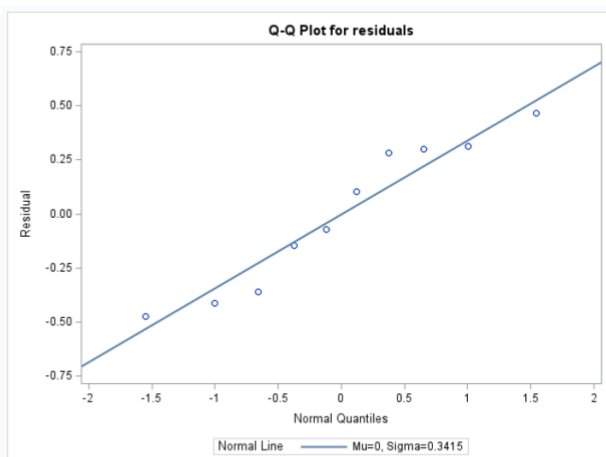
(c) Apply this transformation of Y and generate a scatterplot. Again comment on the appropriateness of using a linear regression model.



If the data was suited for linear regression before, then after the transformation, it should be even more suited now.

(d) Run the regression model using the transformed data and generate a residual plot (using X or $\hat{Y}$) and a normal probability plot. What do the plots show?





The residuals appear to be randomly distributed with constant variance, as desired. Though the QQ plot show clear deviations from the 45-degree line, this can be most likely be explained by the small sample size.

(e) Express the estimated regression function in the original units.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	10.26093	0.21290	48.20	<.0001
year	1	1.07629	0.03988	26.99	<.0001

We know $\sqrt{\hat{Y}^2} = b_0 + b_1X = 10.26093 + 1.07629X$. Then $\hat{Y} = (10.26093 + 1.07629X)^2 = 105.2867 + 22.0875X + 1.1584X^2$.