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

(a) Describe the distribution of the explanatory variable.

A graph of a normal distribution

Description automatically generated

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.

A screenshot of a computer

Description automatically generated

(c) Plot the residuals versus the explanatory variable and briefly describe the plot noting

any unusual patterns or points.

A graph with blue dots

Description automatically generated



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?

A graph of a line

Description automatically generatedA diagram of a distribution of residuals

Description automatically generated

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.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Fitted Equation | t-test for slope | Standard Error | p-value | R2 | Estimate |
| **Original Data** |  | -4.10 | 41.57433 | <0.0001 | 0.1703 | 5552112 |
| **Corrected Data** |  | -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 R2 values.

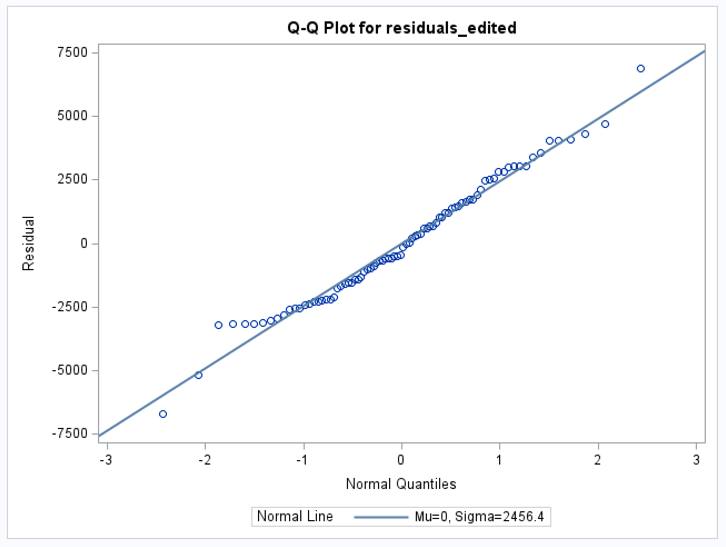
(b) Repeat parts (c) and (d) from the previous problem for this altered data set analysis and summarize how these plots help you to detect the unusual observation.

A graph with blue dots

Description automatically generated



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).

A diagram of a distribution of residuals

Description automatically generated

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.

A graph with numbers and a line

Description automatically generated

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?

A graph with blue lines

Description automatically generated

According to the Box-Cox plot, a 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.

A graph with numbers and a line

Description automatically generated

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 or ) and a normal probability plot. What do the plots show?

A graph with blue dots

Description automatically generatedA graph of a line graph

Description automatically generated

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.

A screenshot of a computer screen

Description automatically generated

We know . Then .