Correlation and Regression

Correlation and Regression

- Correlation and regression both deal with relationships between 2 or more variables.
- ***Correlation***: Estimates the strength of a ***linear*** relationship between variables.
- ***Regression***: Develops an equation that mathematically describes the relationship between two or more the variables.

Correlation

- However, a correlation between two variables does not necessarily imply a cause and effect relationship exists.
  - A strong correlation between two variables does not mean that changing one induces a change in the other. They may both be responding to the effects of one or more other variables that you don't know about.
  - Cause and effect can be determined by a **design of experiment** trial.

Correlation

- The *correlation coefficient* is a numeric measure of the strength of the linear relationship between two variables.
- It is symbolized by ***r***. Minitab calls it the *Pearson Correlation* (for quantitative data).
- ***r*** measures only the strength of a STRAIGHT LINE or linear relationship.
- The correlation coefficient (***r***) can range in value from **-1** to **+1**.

Correlation

- The larger the correlation coefficient, the stronger the linear relationship between the variables.

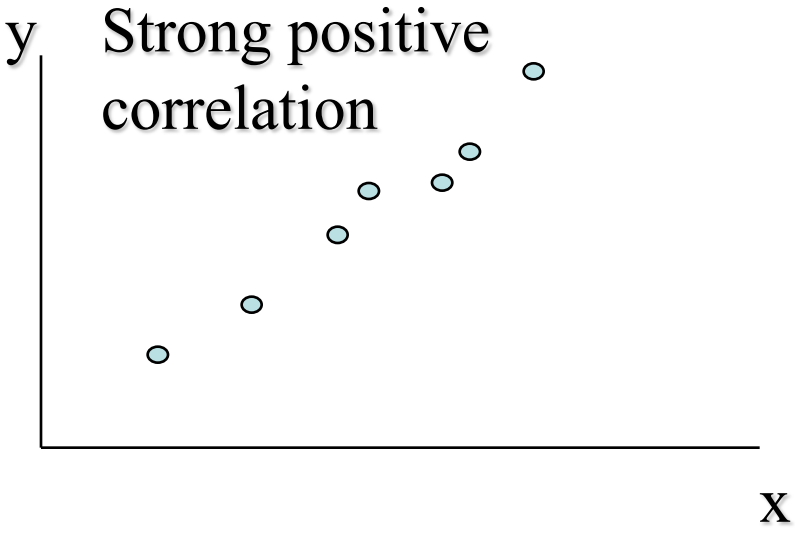
  r = 0 → no correlation

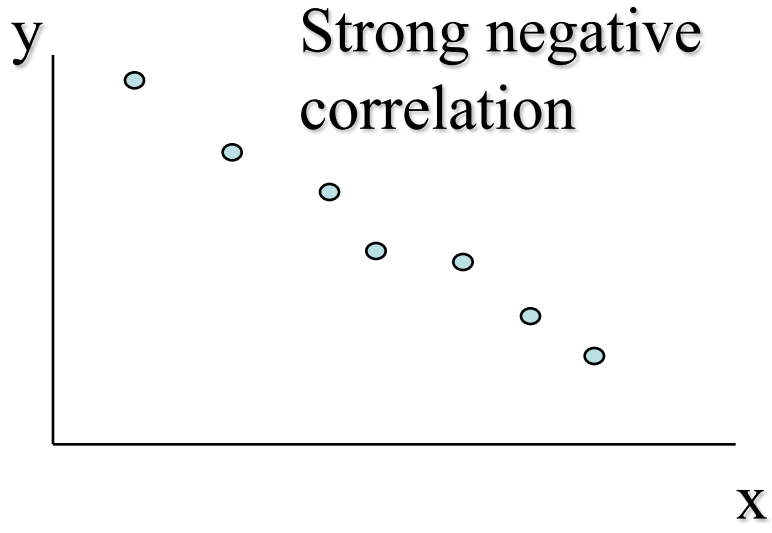
  r = 1 → perfect positive correlation

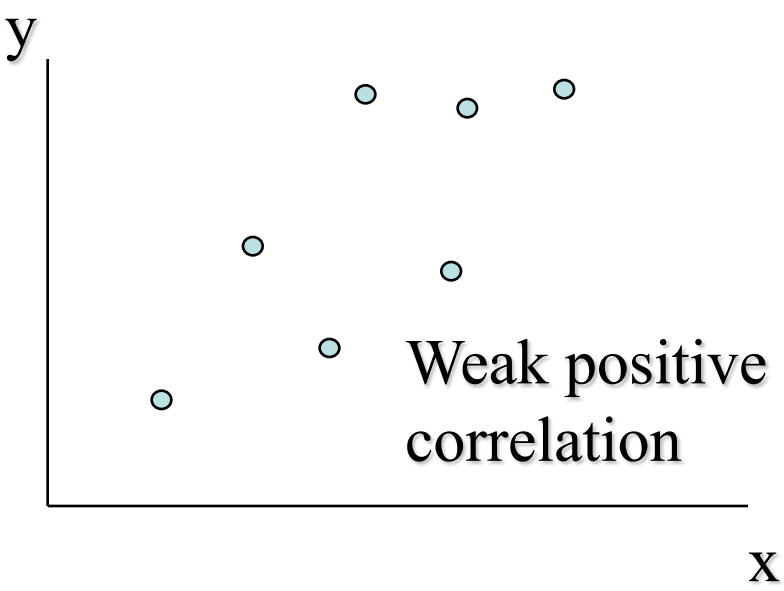
  r = -1 → perfect negative correlation

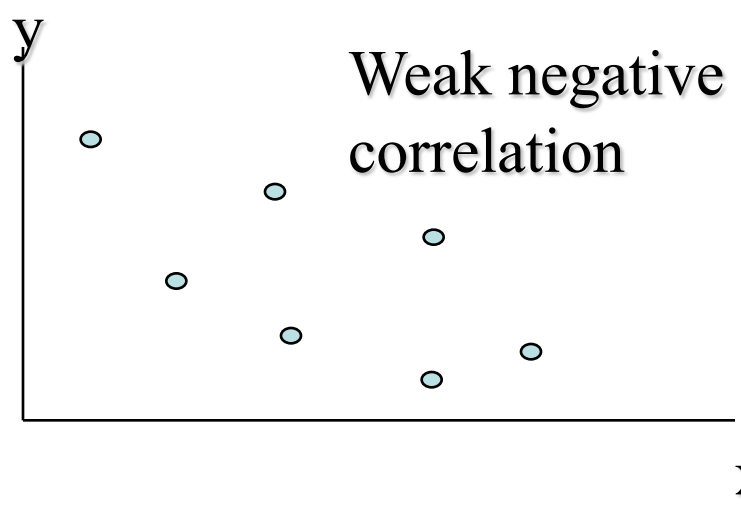
- ***The sign of the coefficient indicates the direction of the relationship.***
  - ***Positive correlation***: If both variables tend to increase or decrease together. $(0 < r \leq 1)$
  - ***Negative correlation***: If one variable tends to increase as the other decreases.$(-1 \leq r < 0)$

Correlation

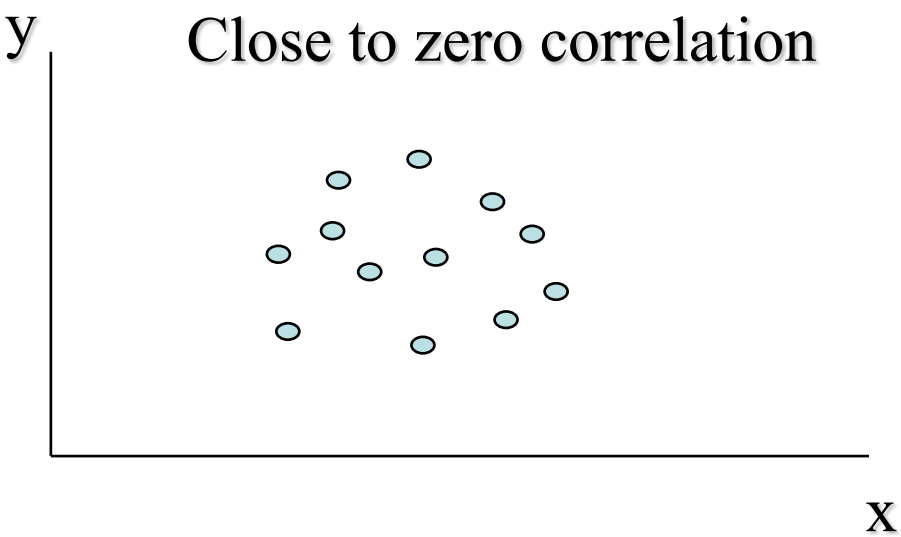


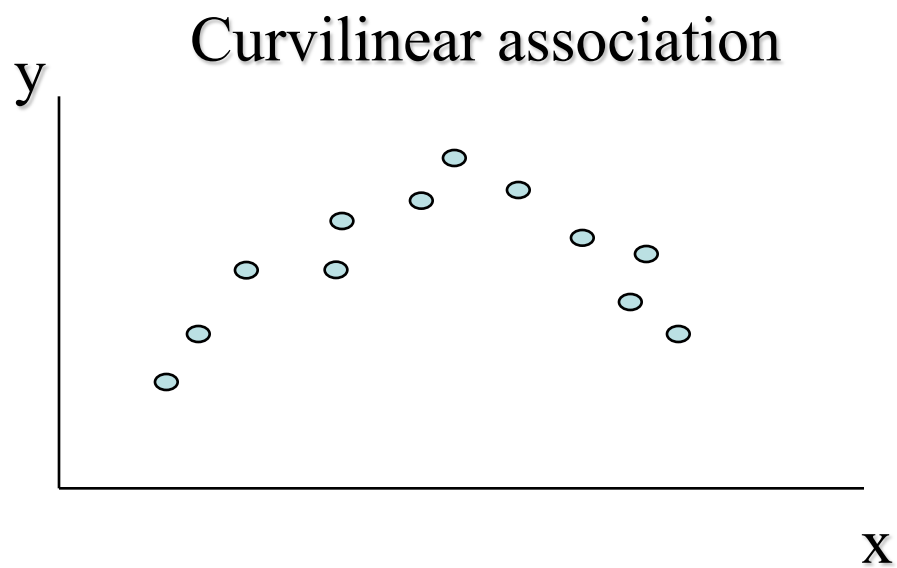






Correlation



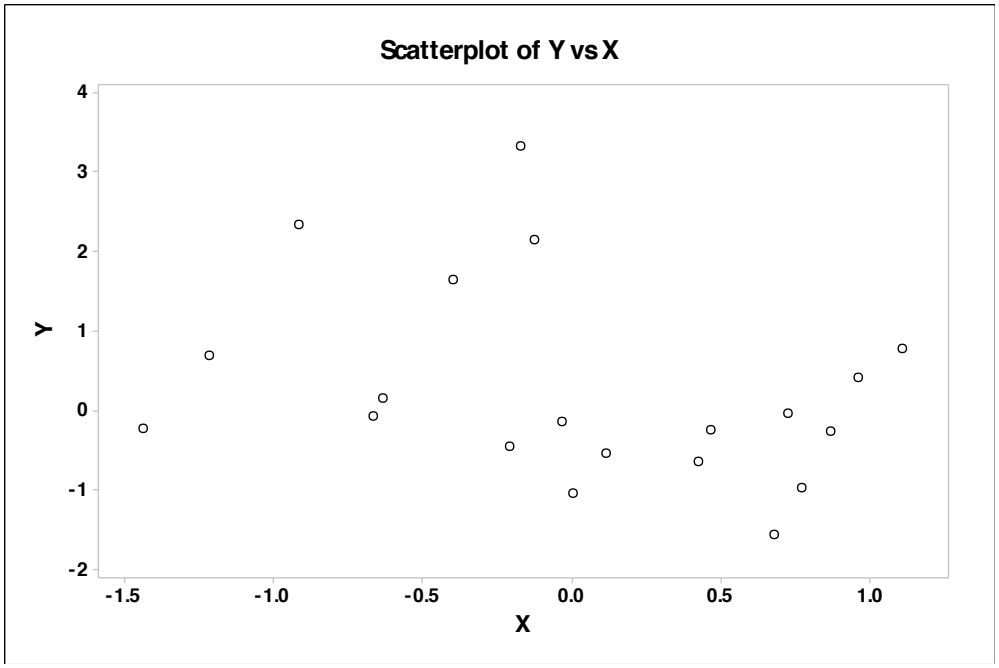


Correlation

- The first step in a **correlation** analysis is to create a scatter plot
- Then calculate the **correlation** coefficient if there is some visual evidence of an ***association***.

Correlation

- A **scatter plot** is used to visualize the relationship or correlation between two variables.

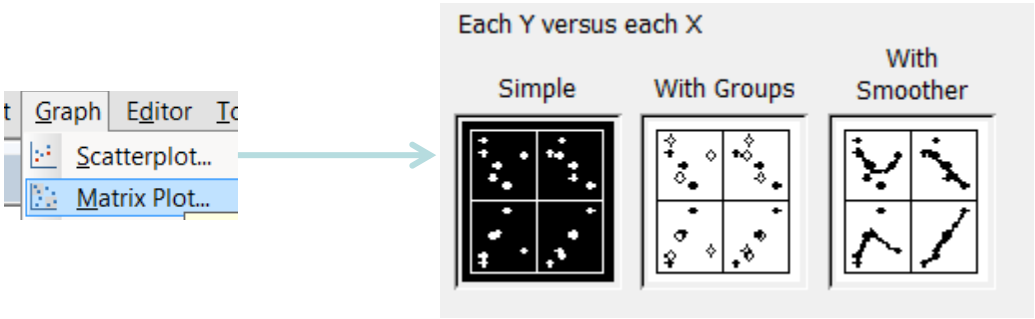


Correlation

- For example, suppose a production manager is interested in finding if there is any relationship to mill natural ***gas use*** and several production measures.

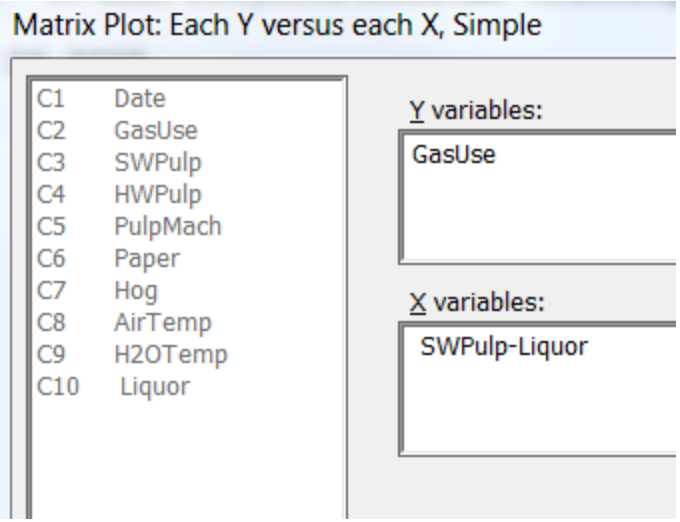
Correlation

- Restart Minitab.
- Open Worksheet ***Minitab Dataset 9.xls***.
- Make a matrix plot of ***GasUse*** through ***Liquor***.
- This time, we'll create a **Simple Each Y versus each X** matrix plot.



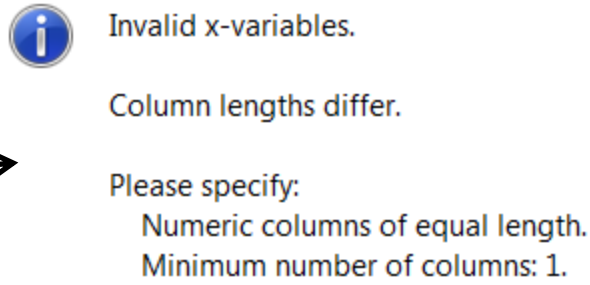
Correlation

- Select **GasUse** as the **Y variable**.
- Select **SWPulp** through **Liquor** as the **X variables**.
- Click **OK**



Correlation

- Do you get this <u>error</u> message?



- This is a Minitab quirk that means there a **blank** in the **last row** of some column.
- Click **OK** → click **Cancel**

Last Value in Column Blank

- Scroll down to the bottom of the worksheet and find the blank in column ***HWPulp***.
- Click in the blank column and type ***** (an asterisk) and hit **Enter** key.
- This will fix the error.

	C4	
lp	HWPulp	Pu
*	1057.0	
*	1024.0	
*	1053.0	
*	1063.0	
2.0	94.0	
5.0		

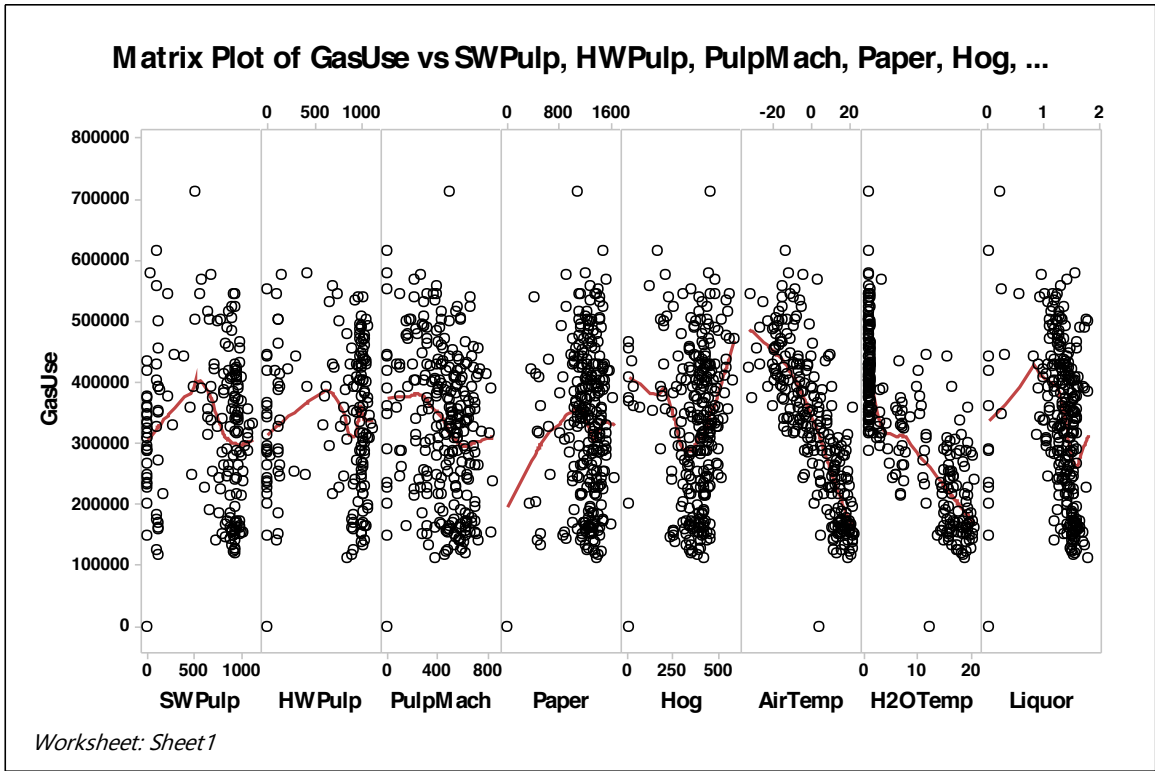
C4
HWPulp
1057.0
1024.0
1053.0
1063.0
94.0
*

Correlation

- Now try to make the **Matrix** Plot **Simple Each Y versus each X** of ***GasUse*** and ***SWPulp*** through ***Liquor*** variables.
- Add a Lowess smoother to the plot.

Correlation

- See any associations with ***GasUse***?



Correlation

- There may be an association between ***GasUse,*** and ***AirTemp*** and ***H2OTemp***.
- To calculate the ***correlation coefficient*** (***r***) between these variables select **Stat** → **Basic Statistics** → **Correlation…** → select ***GasUse***, ***AirTemp*** and ***H2OTemp*** as **Variables** → click **OK**.

Correlation

Correlation: GasUse, AirTemp, H2OTemp, Liquor

```
           GasUse  AirTemp
AirTemp    -0.796
            0.000

H2OTemp    -0.812    0.888
            0.000    0.000

Cell Contents: Pearson correlation
               P-Value
```

r – the correlation coefficient.

p-value (less than .05 indicates ***r*** significantly different from zero).

Correlation

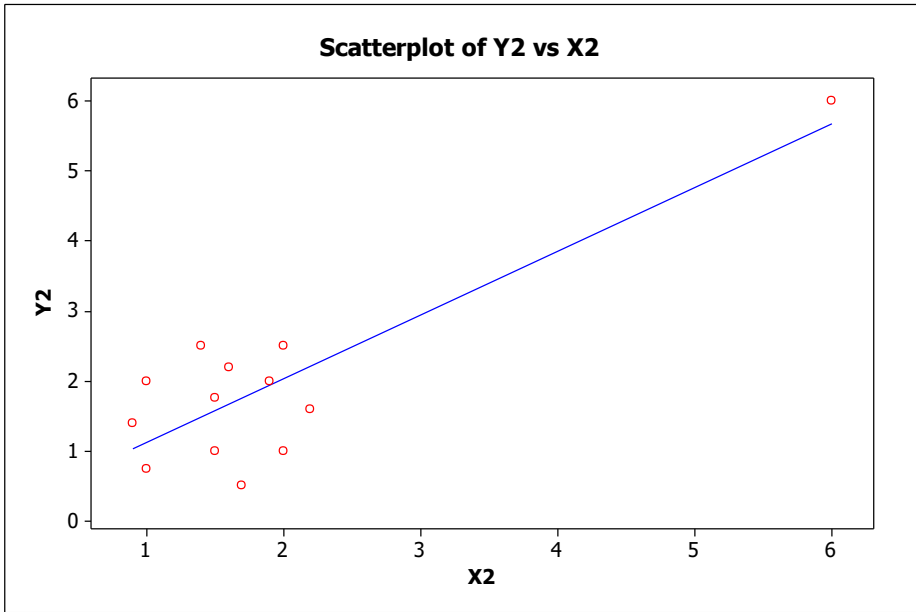
- Interpretation of correlation analysis of ***GasUse,*** and ***AirTemp*** and ***H2OTemp***:

  GasUse is negatively correlated with ***AirTemp*** (***r*** = *-0.796*).

  GasUse is negatively correlated with ***H2OTemp*** (***r*** = *-0.812*).

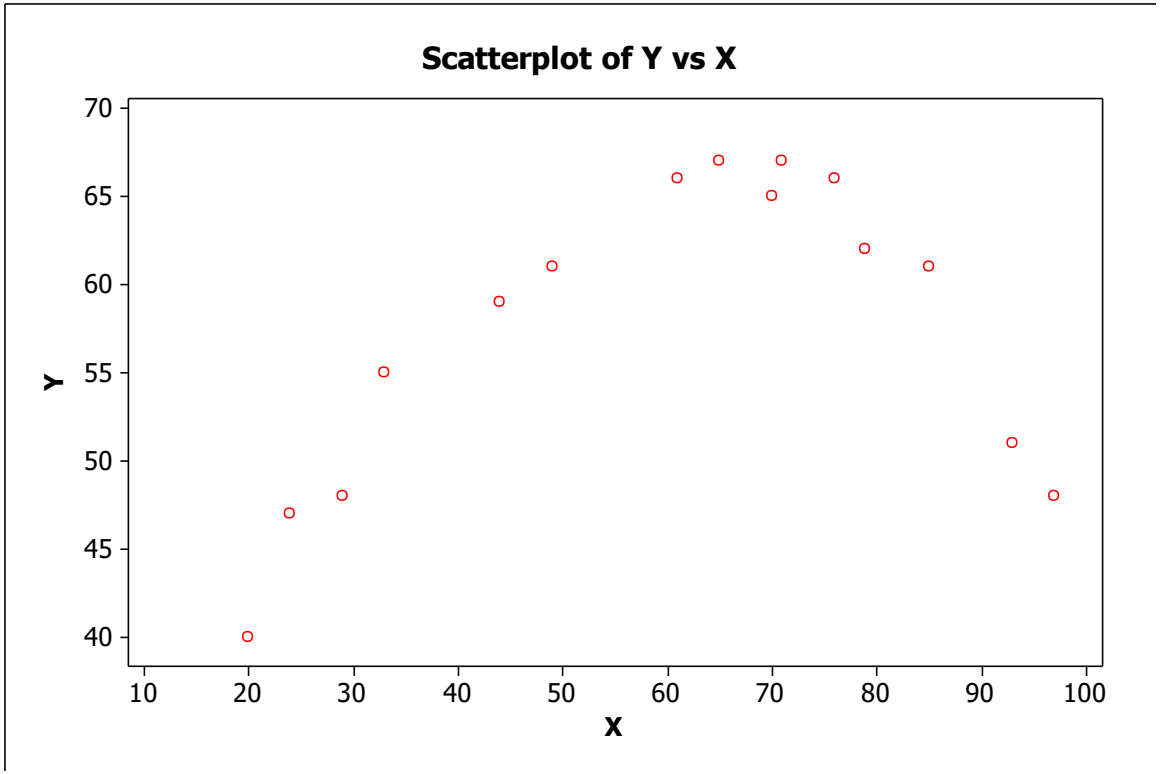
Correlation

- Words of wisdom for **correlation**.
- Just looking at the **correlation** without a scatter plot can lead to *nonsense* results.
- The graph below has a ***r = 0.86*** !



Correlation

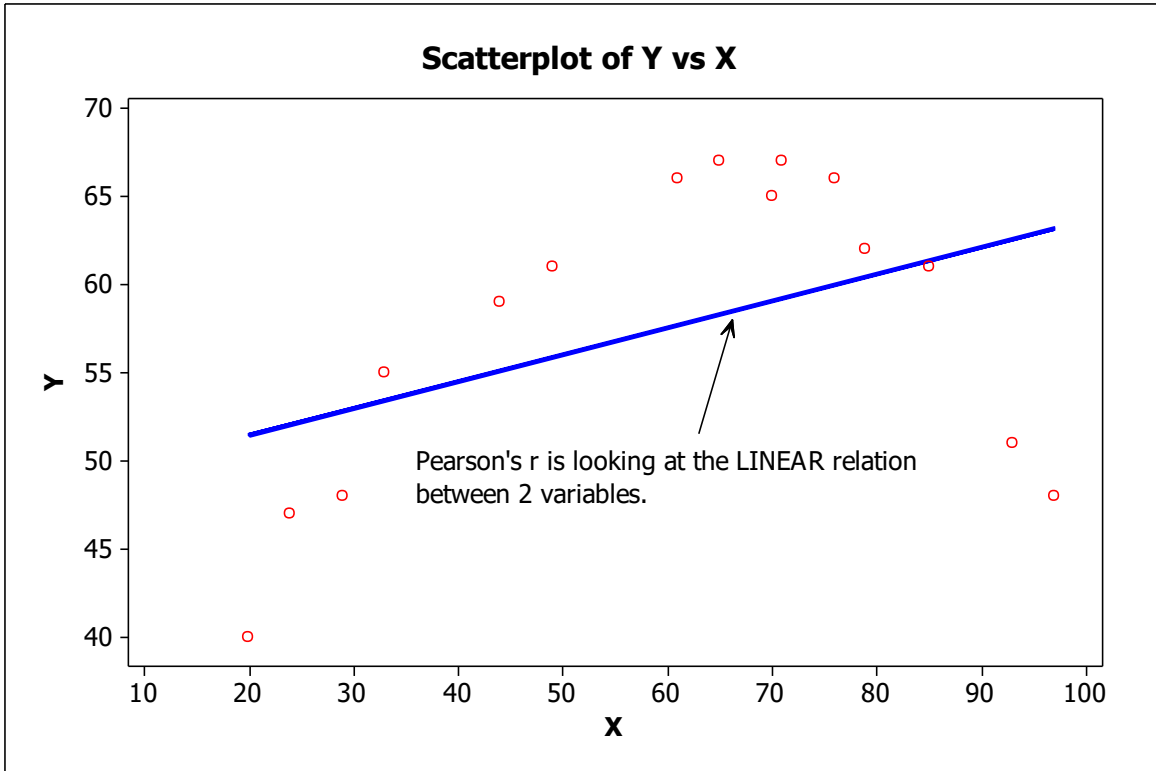
- To emphasize the importance of plotting your data, lets look at a **curved** plot.



Correlation

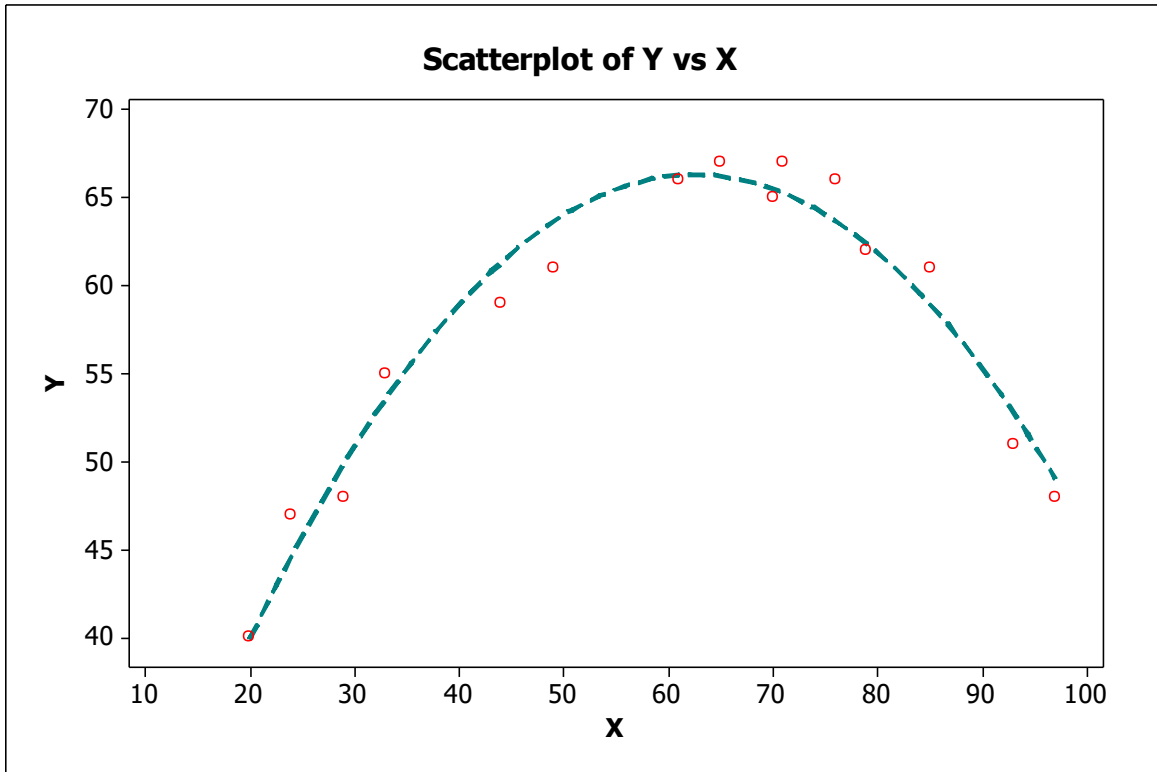
- The result of the correlation analysis is:

  r = 0.44



Correlation

- Clearly the relation between X and Y is **curved** and should be modeled as such.



Simple Regression

- Scatter plots, matrix plots and correlation analysis are all pre-requisites to regression analysis.
- Box plots, histograms and even time series plots are also useful.
- We will start with simple linear regression.

Simple Regression

- **Simple Linear Regression is used when there are only two variables and the relationship is linear (a straight line).**
- The **mathematical model** is assumed to be:

$$y = b_0 + b_1 x + \varepsilon$$

- Similar to the slope-intercept equation of a line.

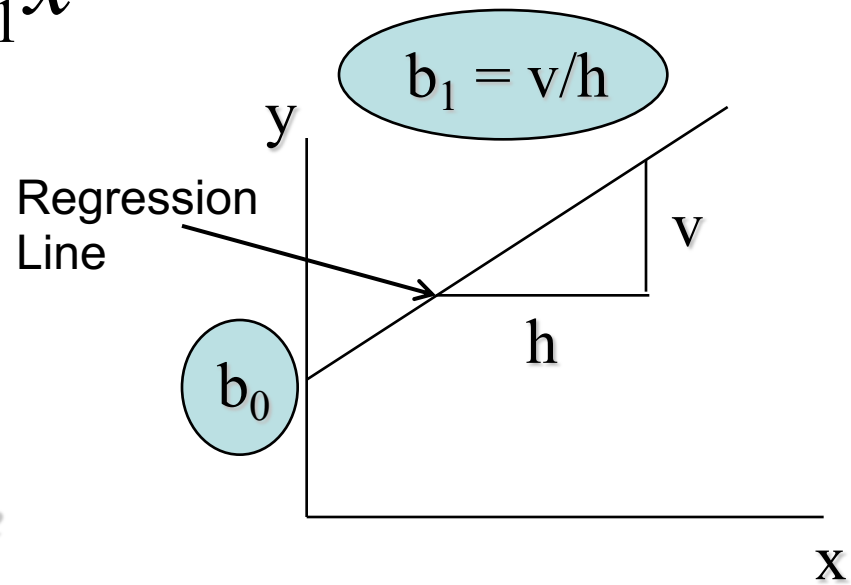
Simple Regression

The Simple Linear Regression Equation

$$y = b_0 + b_1 x$$

this is the *intercept*

this is the *slope*



The regression equation is an algebraic representation of the regression line. It is used to describe the relationship between the response (***y***) and predictor (***x***) variable.

Simple Regression

$$y = b_0 + b_1 x + \varepsilon$$

- The Method of Least Squares is used to determine the *coefficients*, $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$, that define the straight line that "**best**" fits the data.
  - ε (the Greek letter epsilon) is the error.

Simple Regression

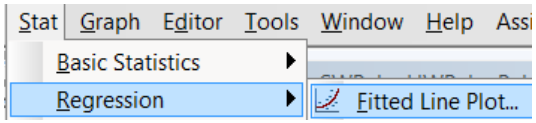
- We will continue to use the worksheet ***Minitab Dataset 9.xls*** for the regression analysis.
- The correlation analysis of ***GasUse***, ***AirTemp*** and ***H2OTemp*** showed that these measures are strongly correlated.

Simple Regression

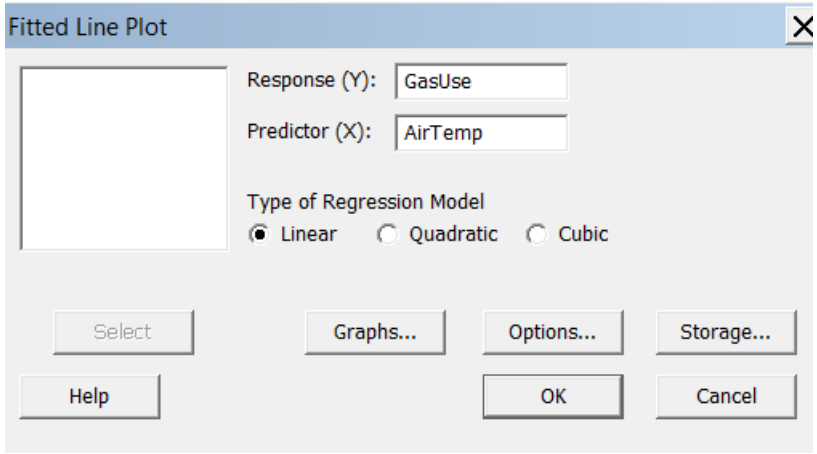
- We will examine the relationship between ***GasUse*** and ***AirTemp***.
  - The ***H2OTemp*** was also correlated with **GasUse**, but it seems the temperature probe got stuck in the ice as there are many 0 °C readings.
- Minitab has a great **simple regression** procedure called the **Fitted Line Plot**.

Simple Regression

- Select **Stat → Regression → Fitted Line Plot…**

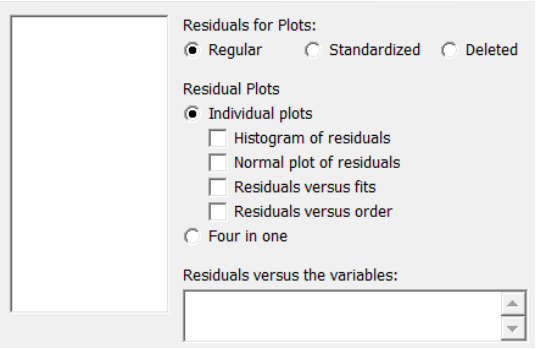


- Select ***GasUse*** as **Response (Y)** and ***AirTemp*** as **Predictor (X).**



Simple Regression

- Leave the **Type of Regression Model** set at the default of **Linear.**
- Click on the **Graphs…** button.
- We are going to learn about the ***Residuals for Plots*** section.



Simple Regression

- The **residual** (ε_i) is the **observed** data value (y_i) minus the models **predicted** value ($\hat{y}_i$) for that observation.

residual = ***observed*** – ***predicted***

- The **residual** value in a regression analysis using the method of least squares is,

$$e_i = y_i - \hat{y}_i$$

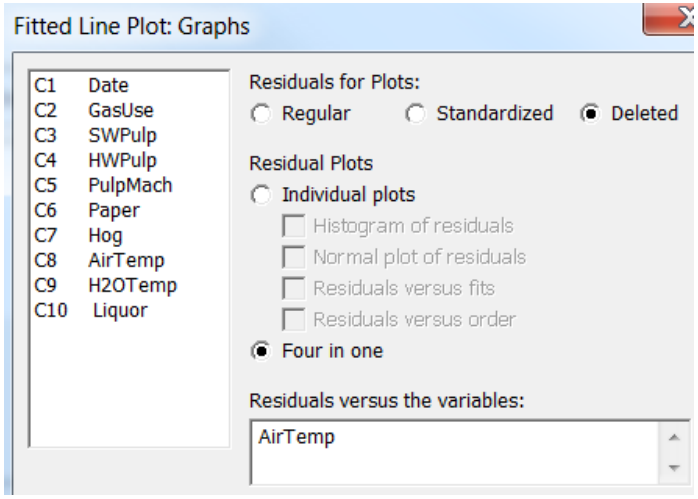
Simple Regression

- There are several residual statistics that have been developed for model diagnostics.
- The one preferred by most statisticians is the deleted (or rstudentized) residual.
- A deleted residual is on a standard deviation scale and is useful for examining the regression analysis model.

Simple Regression

- The **Graphs** dialog is loaded with useful options for regression model diagnostics.
- Choose **Deleted** for the **Residuals for Plots** option.
  - These are the deleted residuals that are scaled to standard deviation units.
- Choose the **Four in one** option.

Simple Regression

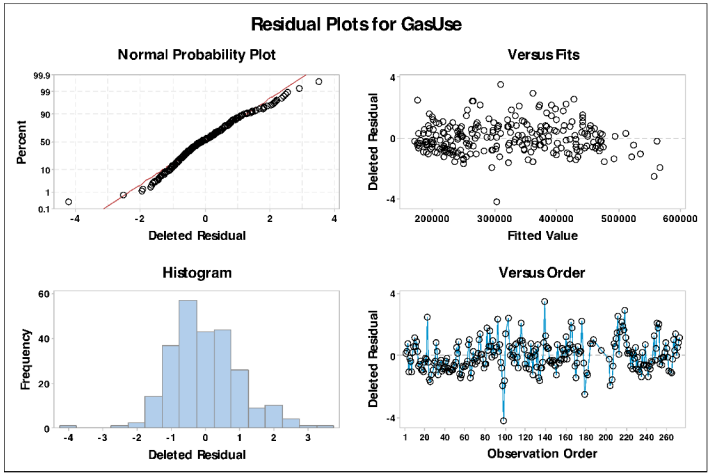


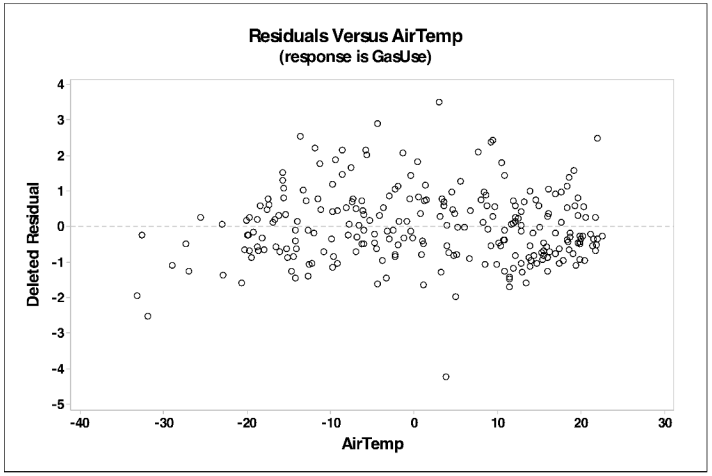
The **Residual versus the variables plots** are useful to check for any unusual patterns across the range of values being measured.

Select ***AirTemp*** and click **OK twice**.

Simple Regression

- Before we look at all of the regression analysis output, let's review some "rules of thumb" that I use when interpreting the associated statistical graphics .



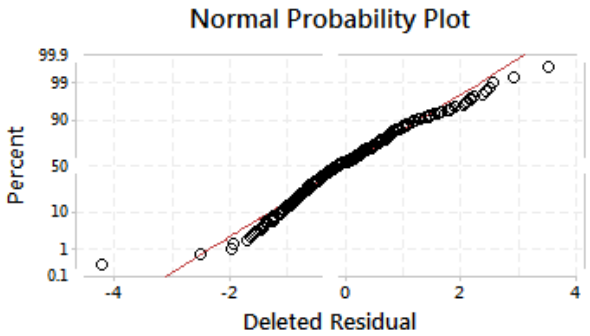


Simple Regression

Residual Plots

- A "rule of thumb" is that a ***deleted residual*** should be ***less than +/-3.***
  - A deleted residual greater than +/-3 may indicate a potential outlier in the data.

Simple Regression

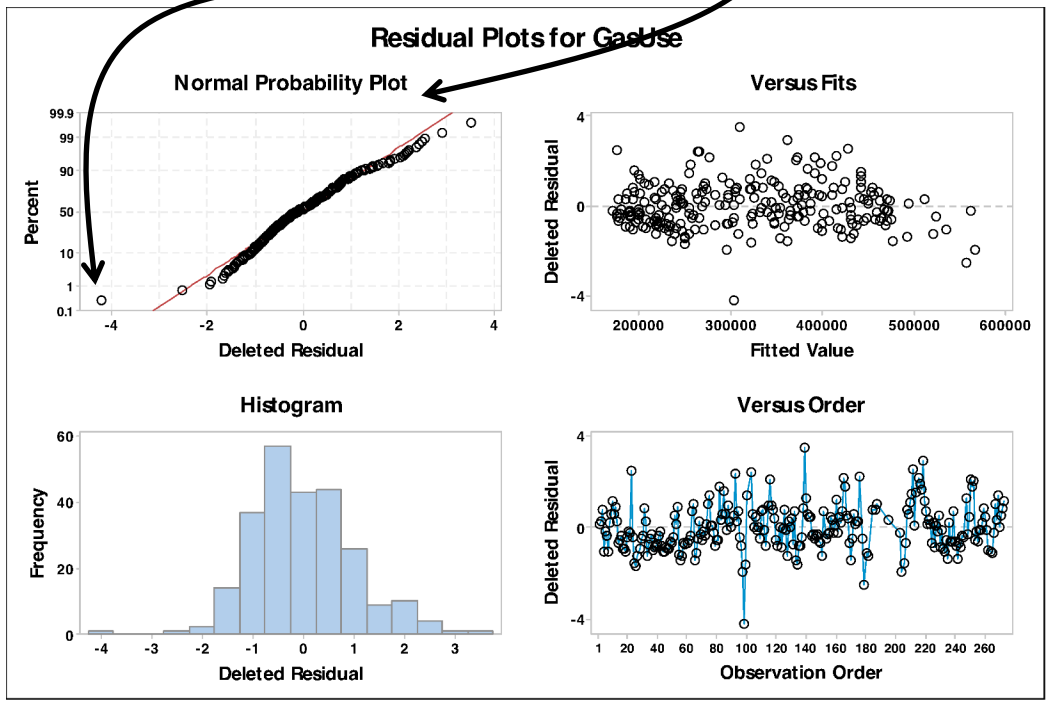


<u>Normal probability plot of residuals</u>

- What you look for are the points ***"hugging"*** the straight line.
- If the deleted residual points departs severely from the Normal line (red), this indicates the linear model may be inappropriate.

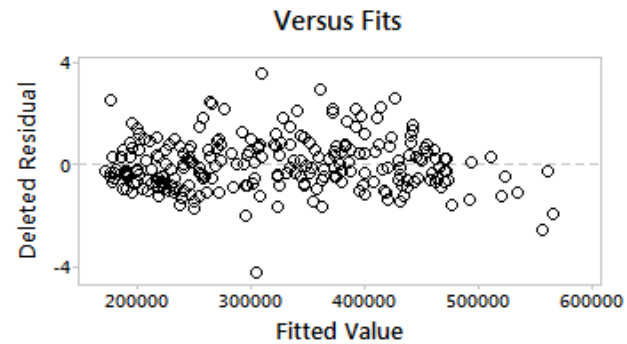
Simple Regression

- Interpretation of normal probability plot: Looks okay, hugging the line.
- There is a potential outlier.

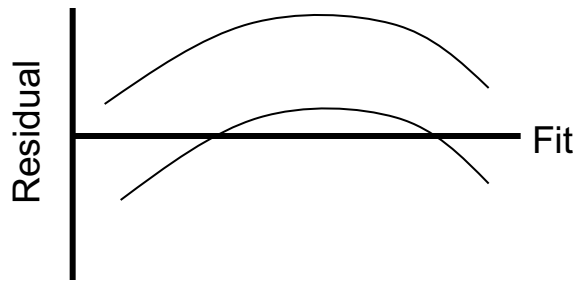


Simple Regression

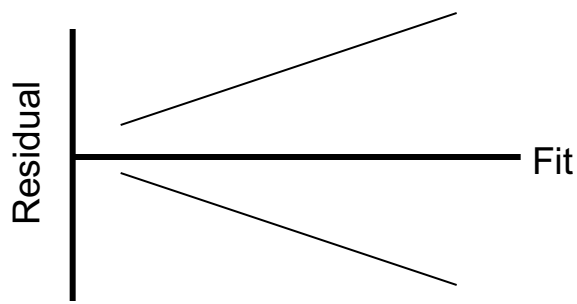
- With the deleted **residuals vs fitted values** plot, you are looking for the following distinct patterns:



Need quadratic term

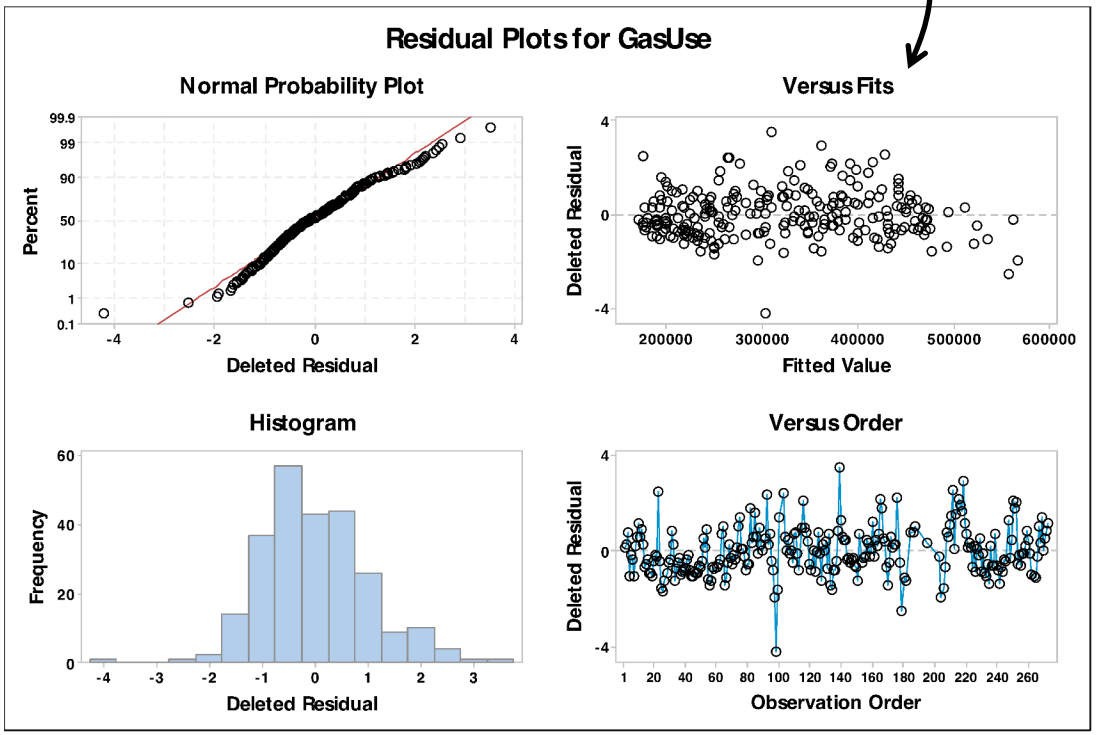


Unequal variance, Need transformation



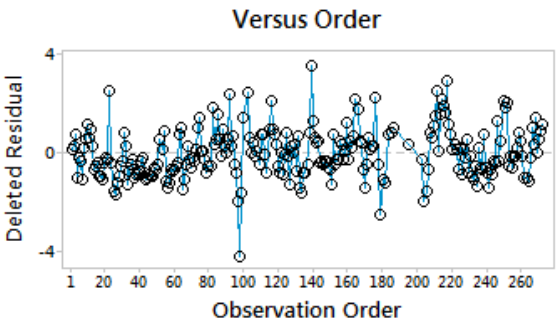
Simple Regression

- Interpretation of Residuals vs Fits plot: Looks okay, random looking.

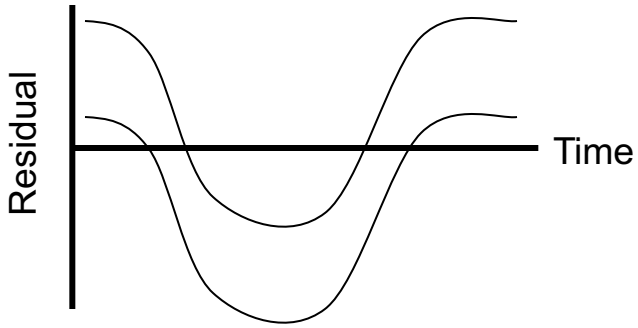


Simple Regression

- The deleted <u>residuals vs time order</u> plot is used to detect a systematic shift in the data over time.
  - Example below of trending over time.

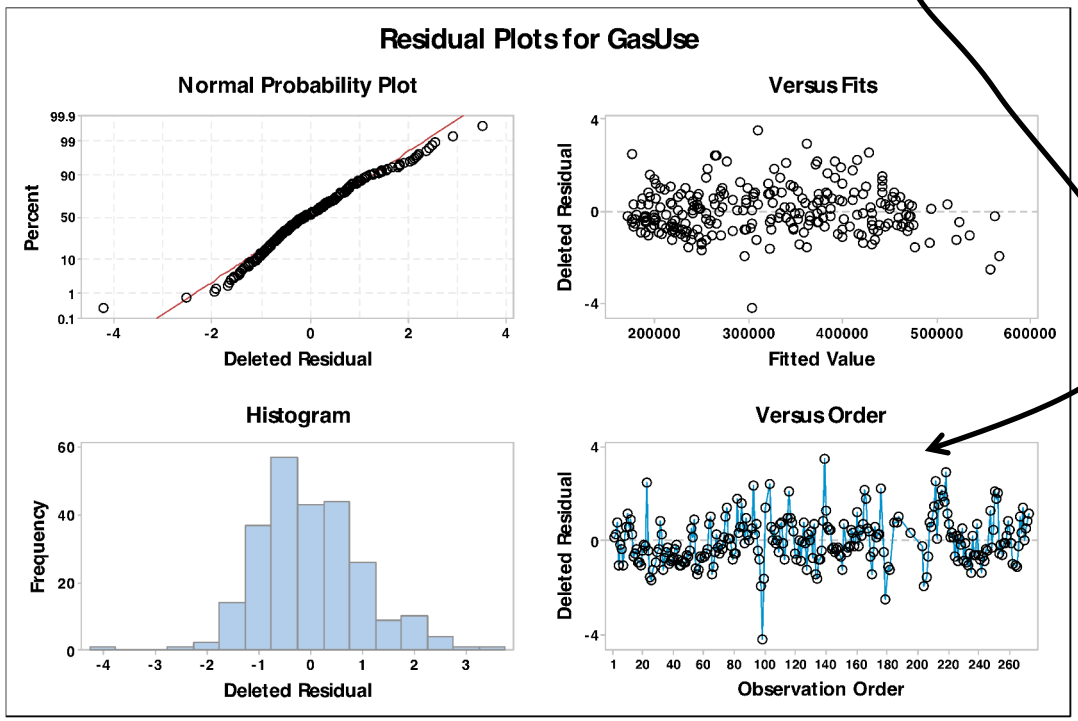


Time trend in data



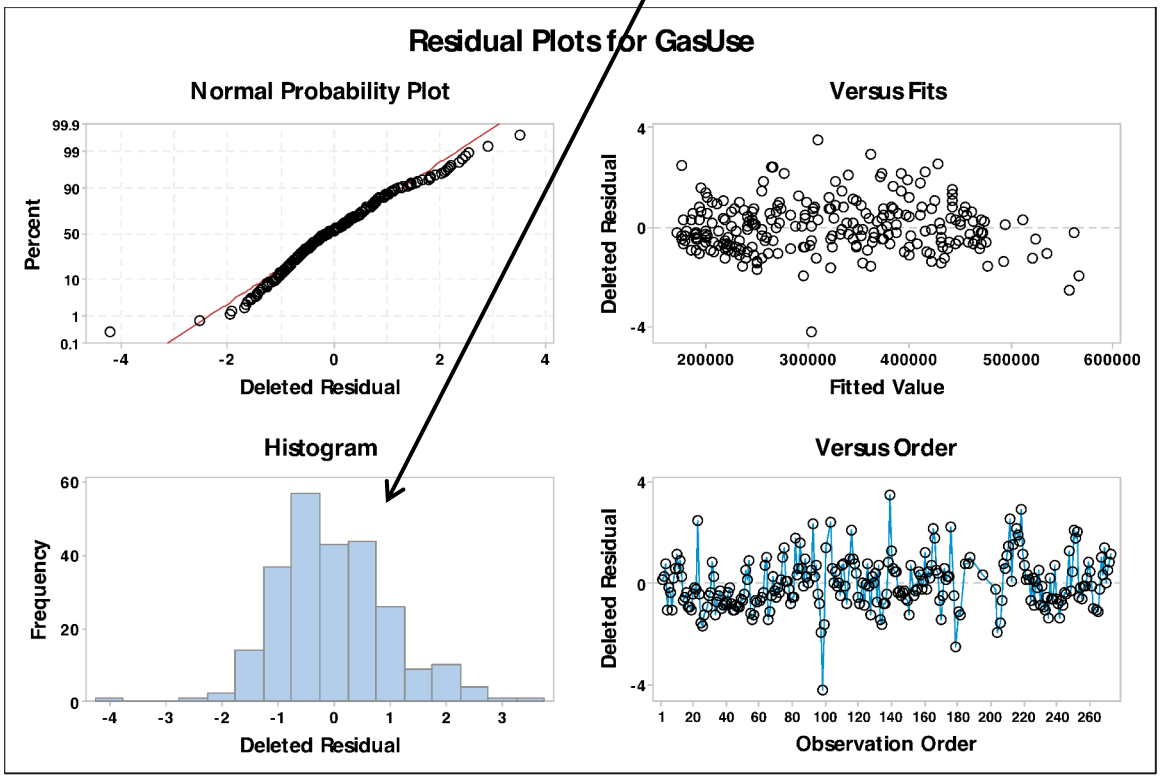
Simple Regression

- Interpretation of <u>Residuals vs Order</u>: Looks okay, but maybe slight time trend?



Simple Regression

- I don't look at the <u>Histogram of the Residuals</u> too hard. If it's "mound-shaped" that's good enough.

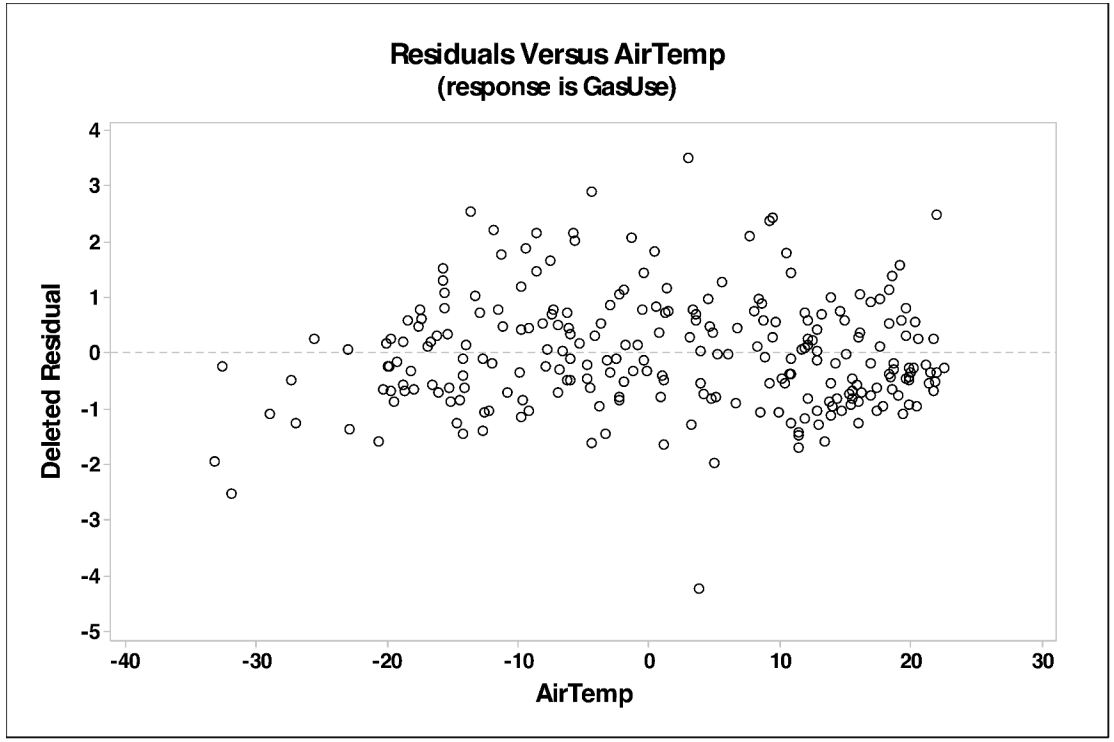


Simple Regression

- The **deleted residuals versus observed variables** should show only randomness.
- We are looking for non-random patterns that would indicate problems.
  - ***Time trends, curved pattern or unequal variance (fan) pattern.***

Simple Regression

- Interpretation of Residuals vs AirTemp plot: Looks okay, random.



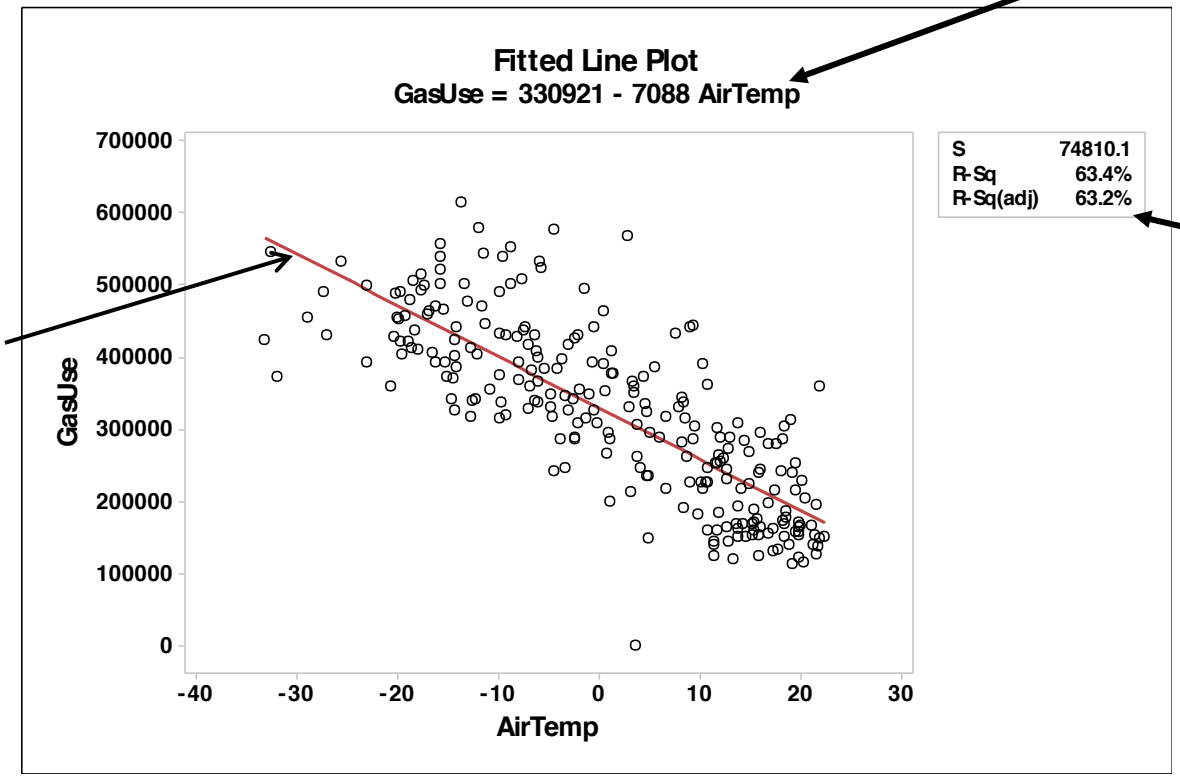
Simple Regression

- Now on to interpreting the regression analysis results.
- The **<u>fitted line plot</u>** display the data points, regression line, regression equation and model fit statistics.

Simple Regression

The fitted line plot.

Regression equation.



Regression line (red).

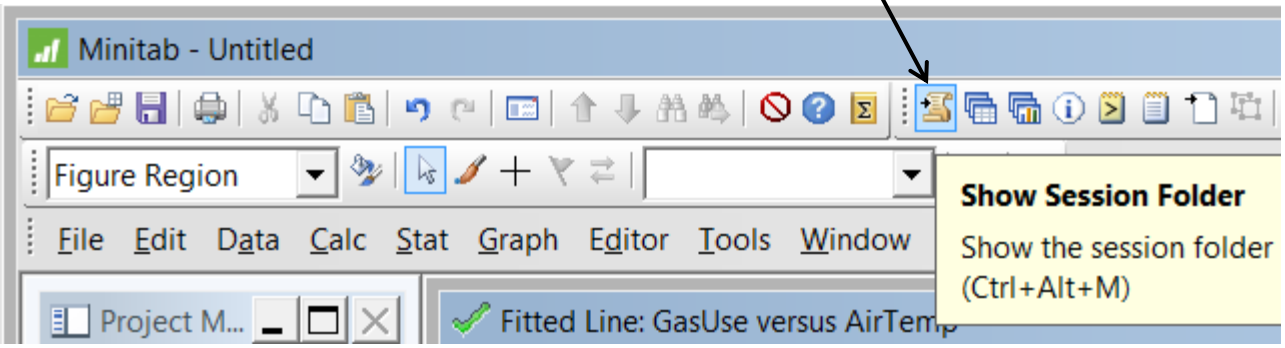
S (RMSE), R-square, Adjusted R-square.

Simple Regression

- R-square
  - R^2 describes the amount of variation in the observed *y* values that is explained by *x*. For simple regression this is the correlation coefficient, *r*, squared.
  - Note that Adj R^2 and R^2 are nearly equal for simple regression. Adj R^2 adjusts for the number of X's in the model.

Simple Regression

- The fitted line plot procedure also gives text output in the **Session** window.
- The easiest way to get to the **Session** window is to click the shortcut icon on the ribbon like we did for the worksheet.



Simple Regression

Regression Analysis: GasUse versus AirTemp

```
The regression equation is
GasUse = 330921 - 7088 AirTemp

S = 74810.1   R-Sq = 63.4%   R-Sq(adj) = 63.2%

Analysis of Variance

Source        DF           SS           MS       F      P
Regression     1  2.40363E+12  2.40363E+12  429.48  0.000
Error        248  1.38794E+12  5.59655E+09
Total        249  3.79157E+12
```

The additional info is the Analysis of Variance table. The key statistic to look at is the **P**-value. If the **P**-value < 0.05, then we say the coefficient is significantly different from 0. The **P**-value = **0.000**, which is less than 0.05, therefore the coefficient is significantly different from 0.

Simple Regression

- Regression Analysis Conclusions:
  - There may be a potential outlier that needs to be investigated.
    - Run the regression analysis without the potential outlier to see if effects the results.
  - There may be seasonality to the data.
    - Use a seasonal adjustment.
  - AirTemp is a fairly good predictor of GasUse.
    - $r^2 = 0.634$

Simple Regression

- Minitab also has a module to do regression analysis.
  - The Fitted Line Plot output is the same for simple regression and I prefer it for the plot itself.
- The regression procedure must be used for multiple regression analysis.
  - multiple regression means more than one ***X*** or ***predictor*** variables.

EXERCISE

- Suppose an engineer was interested in whether or not ***Brightness*** and ***Color L*** are related to each other.
- Use graphing, correlation and simple linear regression to test if a relationship exists.
- The data is in ***Exercise 4.xlsx***.