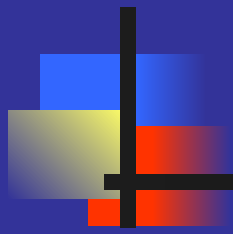
Introduction to Correlation and Regression

Ginger Holmes Rowell, Ph. D.
Associate Professor of Mathematics
Middle Tennessee State University

MIDDLE TENNESSEE STATE UNIVERSITY

Outline

- Introduction
- Linear Correlation
- Regression
  - Simple Linear Regression
  - Using the TI-83
  - Model/Formulas



Outline continued

- Applications
  - Real-life Applications
  - Practice Problems
- Internet Resources
  - Applets
  - Data Sources

Correlation

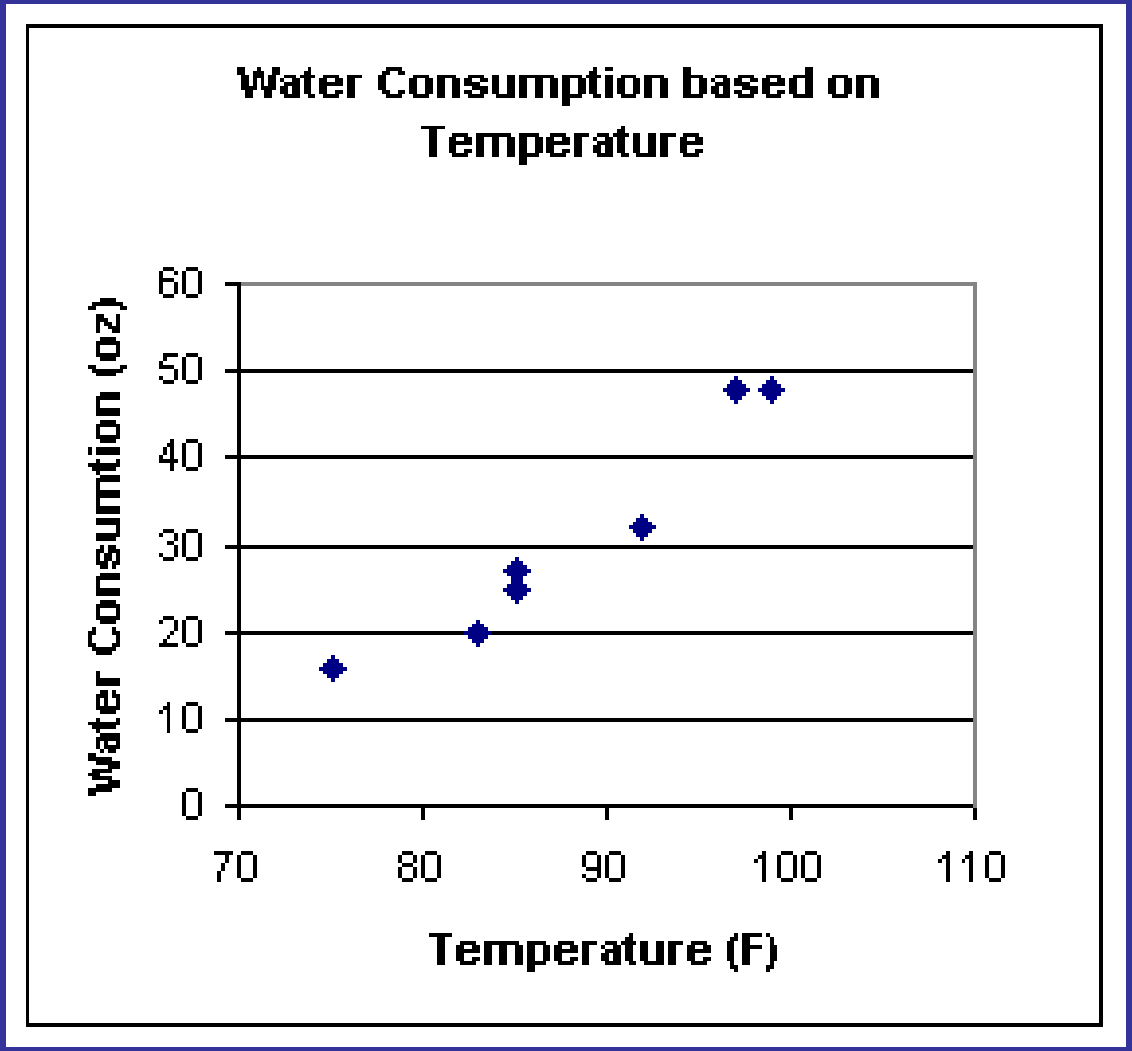
- **Correlation**
  - A measure of association between two numerical variables.
- **Example (positive correlation)**
  - Typically, in the summer as the temperature increases people are thirstier.

Specific Example

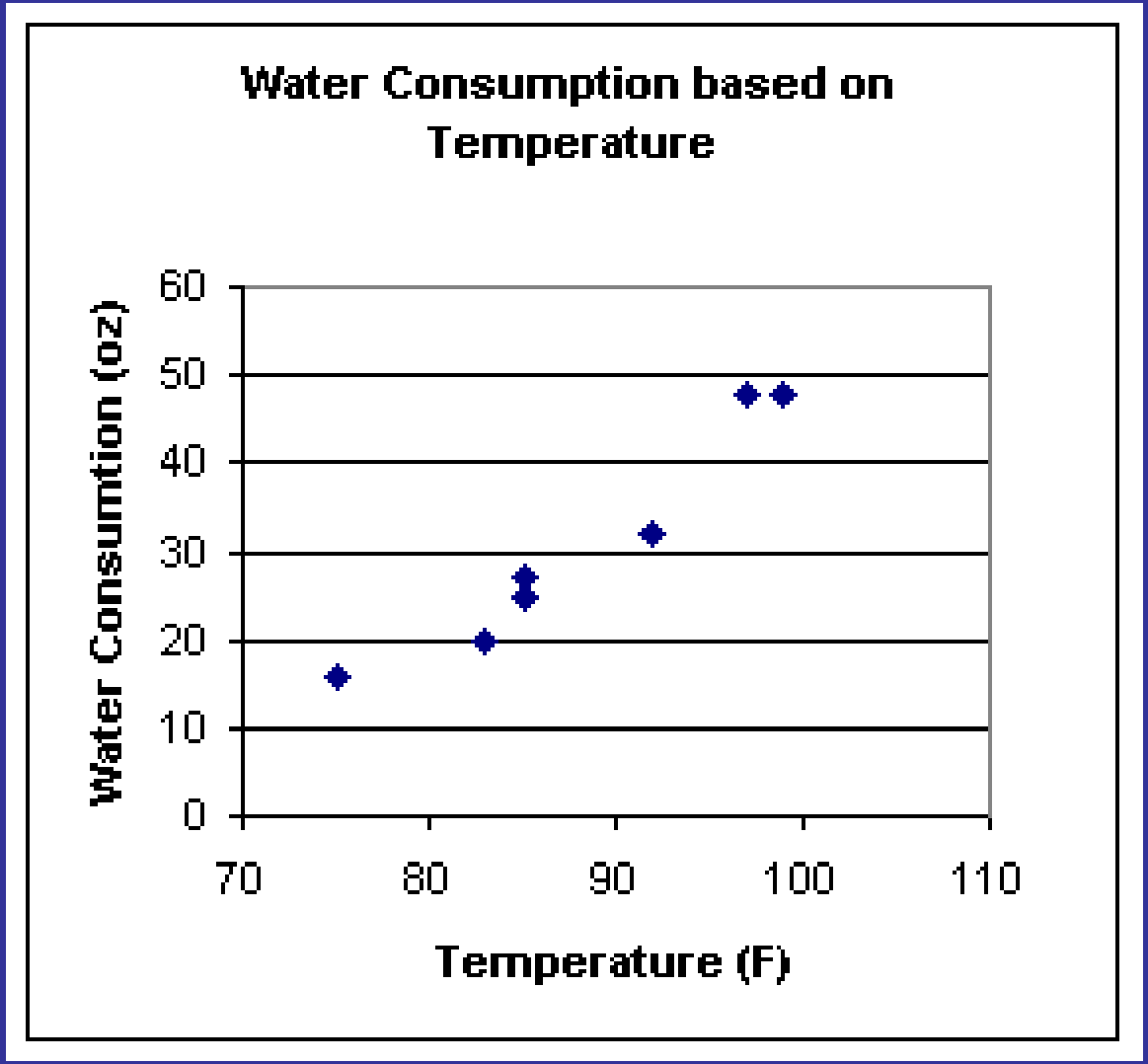
For seven random summer days, a person recorded the temperature and their water consumption, during a three-hour period spent outside.

Temperature (F)	Water Consumption (ounces)
75	16
83	20
85	25
85	27
92	32
97	48
99	48

How would you describe the graph?



How “strong” is the linear relationship?



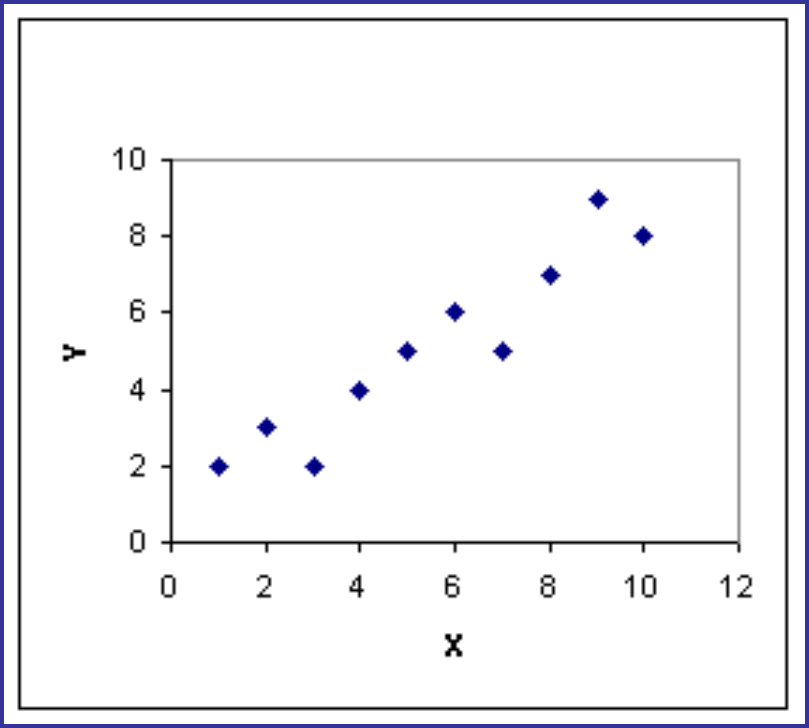
Measuring the Relationship

Pearson's Sample Correlation Coefficient, *r*

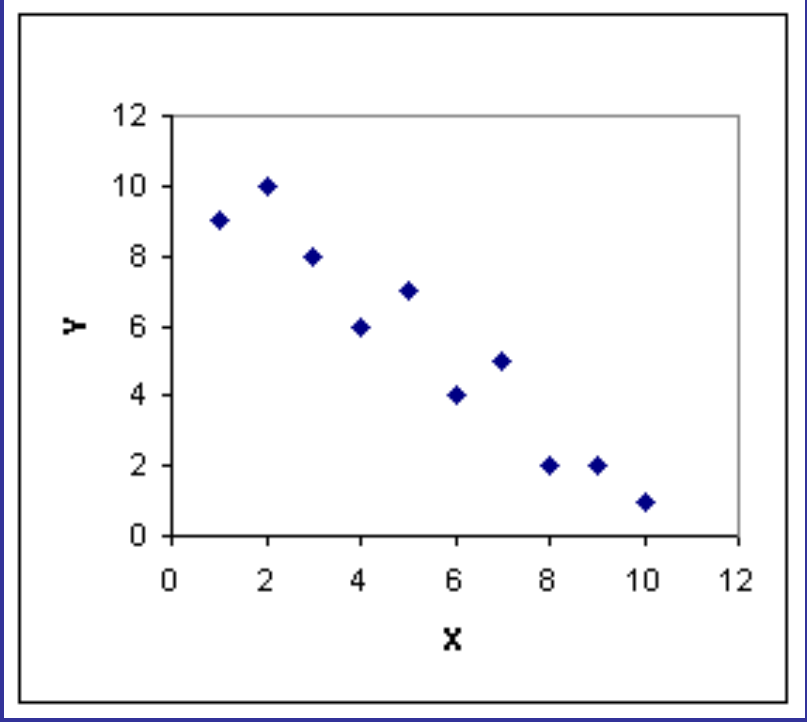
measures the **direction** and the **strength** of the linear association between two numerical paired variables.

Direction of Association

Positive Correlation



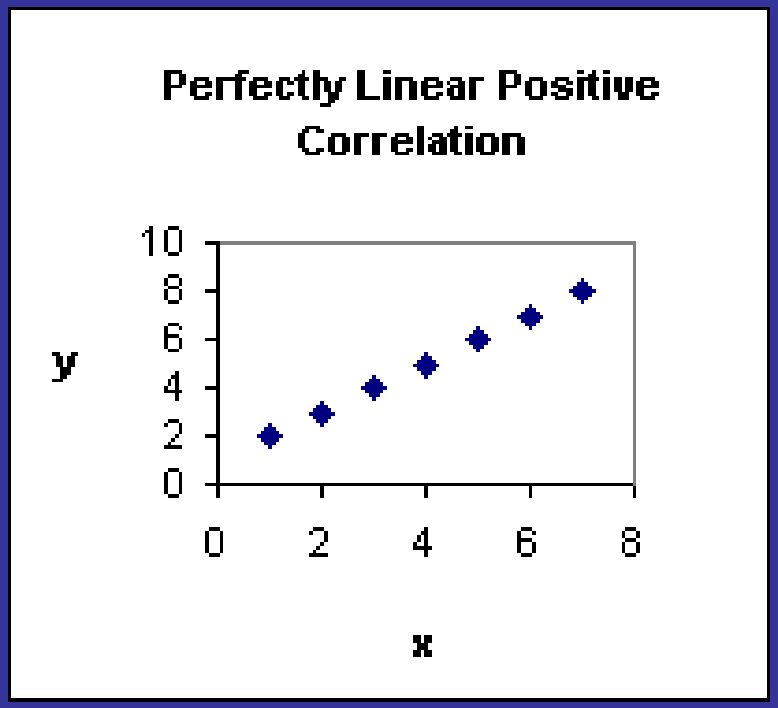
Negative Correlation

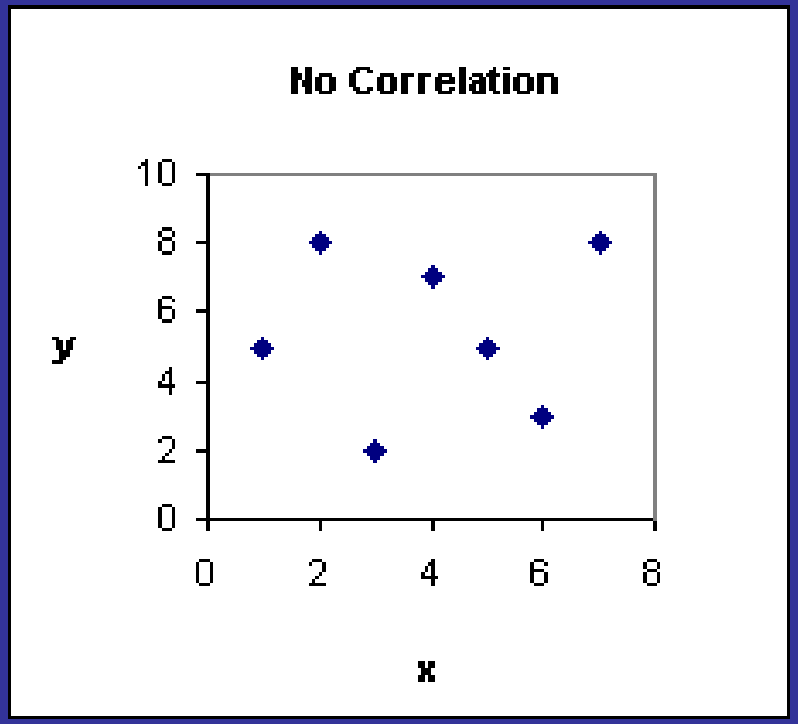


Strength of Linear Association

r value	**Interpretation**
1	perfect positive linear relationship
0	no linear relationship
-1	perfect negative linear relationship

Strength of Linear Association

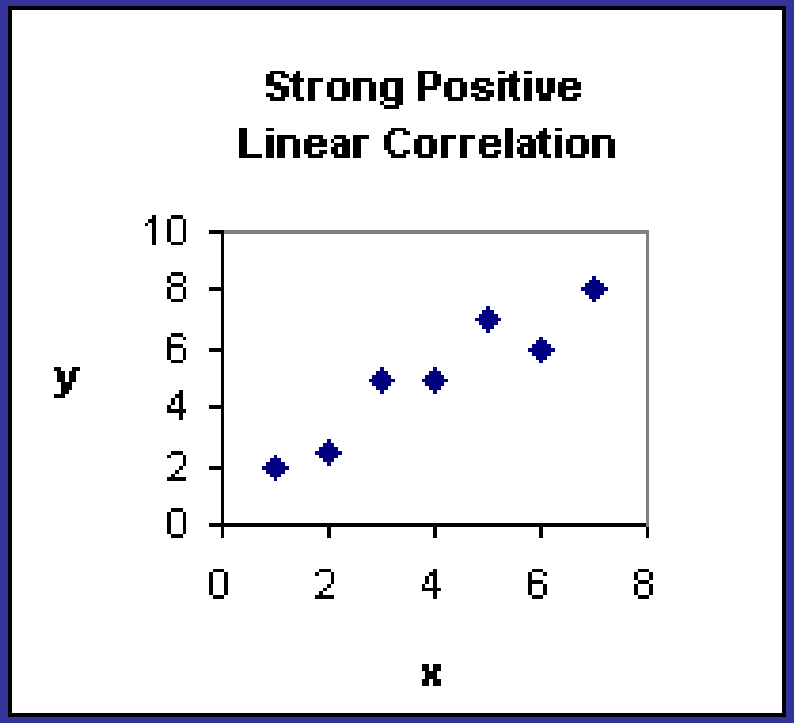


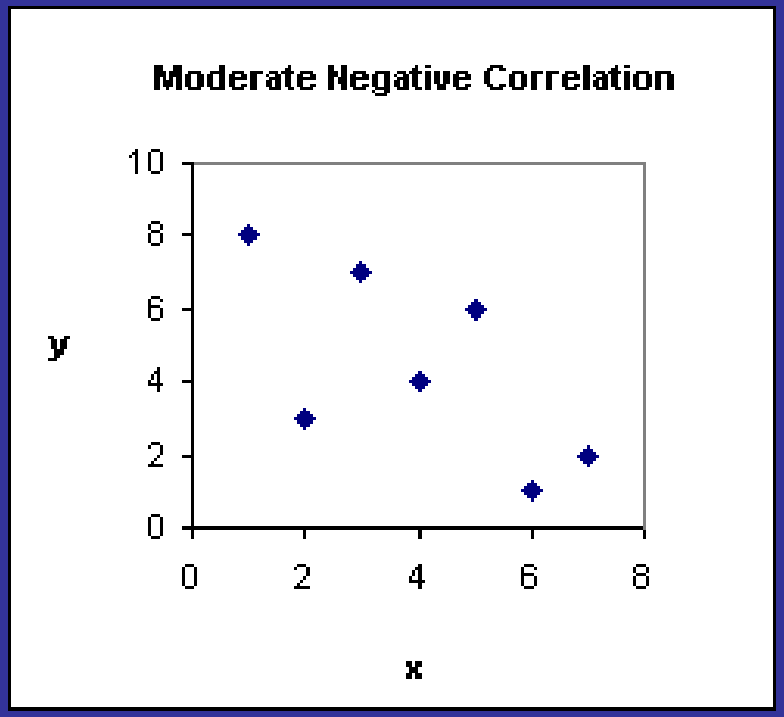


Other Strengths of Association

r value	**Interpretation**
0.9	strong association
0.5	moderate association
0.25	weak association

Other Strengths of Association





Formula

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{s_x} \frac{(y_i - \bar{y})}{s_y}$$

Σ = the sum		
n = number of paired items		
x_i = input variable		y_i = output variable
$\bar{x}$ = x-bar = mean of x's		$\bar{y}$ = y-bar = mean of y's
s_x= standard deviation of x's		s_y= standard deviation of y's

Regression

- **Regression**
  - **Specific statistical methods** for finding the "line of best fit" for one response (dependent) numerical variable based on one or more explanatory (independent) variables.

Curve Fitting vs. Regression

- **Regression**
  - Includes using statistical methods to assess the "goodness of fit" of the model. (ex. Correlation Coefficient)

Regression: 3 Main Purposes

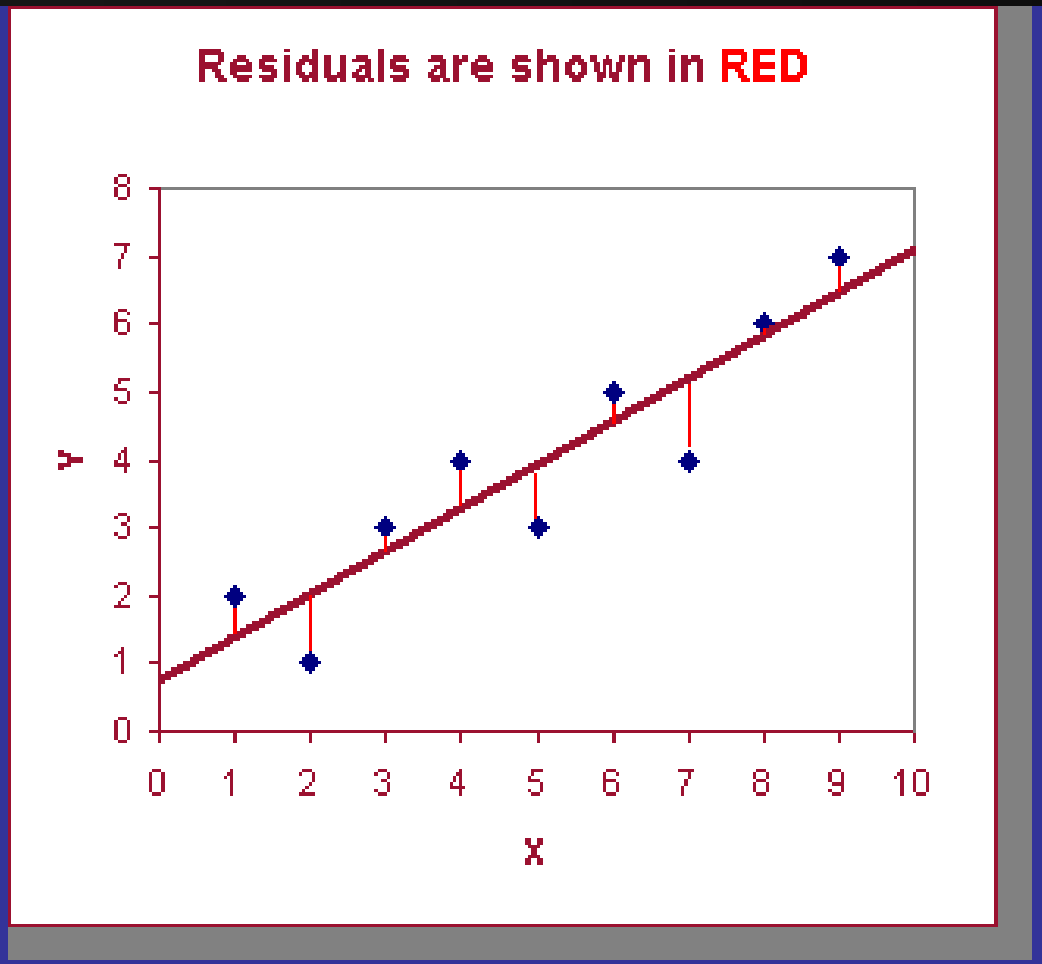
- **To describe** (or model)
- **To predict** (or estimate)
- **To control** (or administer)

Simple Linear Regression

- Statistical method for finding
  - the “line of best fit”
  - for one response (dependent) numerical variable
  - based on one explanatory (independent) variable.

Least Squares Regression

- **GOAL** - minimize the sum of the square of the errors of the data points.



- This minimizes the **Mean Square Error**

Example

- Plan an outdoor party.
- **Estimate** number of soft drinks to buy per person, based on how hot the weather is.
- Use Temperature/Water data and **regression**.

Steps to Reaching a Solution

- Draw a scatterplot of the data.

Steps to Reaching a Solution

- Draw a scatterplot of the data.
- Visually, consider the strength of the linear relationship.

Steps to Reaching a Solution

- Draw a scatterplot of the data.
- Visually, consider the strength of the linear relationship.
- If the relationship appears relatively strong, find the correlation coefficient as a numerical verification.

Steps to Reaching a Solution

- Draw a scatterplot of the data.
- Visually, consider the strength of the linear relationship.
- If the relationship appears relatively strong, find the correlation coefficient as a numerical verification.
- If the correlation is still relatively strong, then find the simple linear regression line.

Our Next Steps

- Learn to Use the TI-83 for Correlation and Regression.
- Interpret the Results (in the Context of the Problem).

Finding the Solution: TI-84

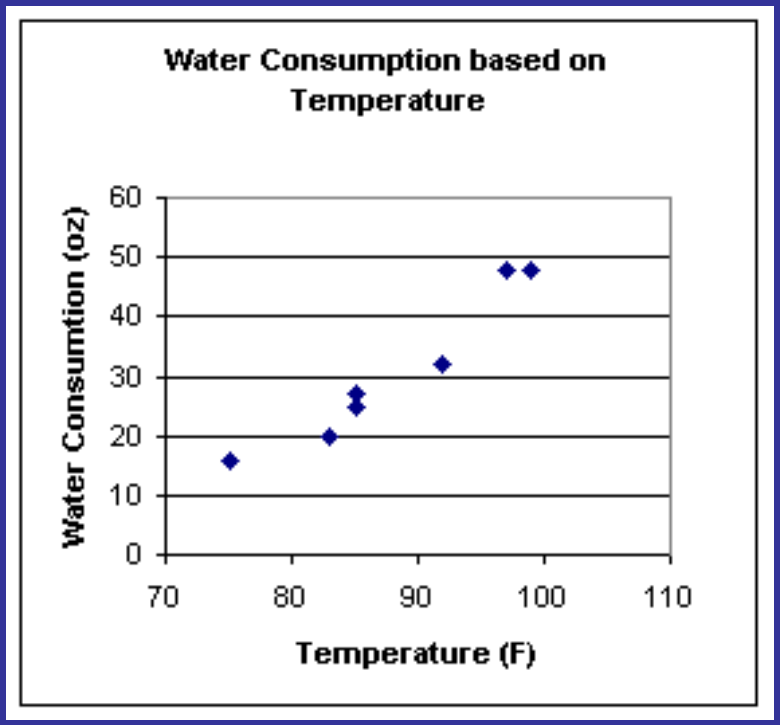
- Using the TI- 83 graphing calculator
  - Turn on the calculator diagnostics.
  - Enter the data.
  - Graph a scatterplot of the data.
  - Find the equation of the regression line and the correlation coefficient.
  - Graph the regression line on a graph with the scatterplot.

Preliminary Step

- Turn the Diagnostics On.
  - Press **2nd 0** (for Catalog).
  - Scroll down to **DiagnosticOn**. The marker points to the right of the words.
  - Press **ENTER**. Press **ENTER** again.
  - The word **Done** should appear on the right hand side of the screen.

Example

Temperature (F)	Water Consumption (ounces)
75	16
83	20
85	25
85	27
92	32
97	48
99	48



1. Enter the Data into Lists

- Press **STAT**.
- Under **EDIT**, select **1: Edit**.
- Enter x-values (input) into **L1**
- Enter y-values (output) into **L2**.
- After data is entered in the lists, go to **2nd MODE** to quit and return to the home screen.
- Note: If you need to clear out a list, for example list 1, place the cursor on L1 then hit CLEAR and ENTER .

2. Set up the Scatterplot.

- Press **2nd Y=** (STAT PLOTS).
- Select **1: PLOT 1** and hit **ENTER**.
- Use the arrow keys to move the cursor down to **On** and hit **ENTER**.
- Arrow down to **Type:** and select the **first graph** under Type.
- Under **Xlist:** Enter **L1**.
- Under **Ylist:** Enter **L2**.
- Under **Mark:** select any of these.

3. View the Scatterplot

- Press **2nd MODE** to quit and return to the home screen.
- To plot the points, press **ZOOM** and select **9: ZoomStat**.
- The scatterplot will then be graphed.

4. Find the regression line.

- Press **STAT**.
- Press **CALC**.
- Select **4: LinReg(ax + b)**.
- Press **2nd 1** (for List 1)
- Press the **comma key**,
- Press **2nd 2** (for List 2)
- Press **ENTER**.

5. Interpreting and Visualizing

- Interpreting the result:

$$y = ax + b$$

  - The value of ***a*** is the **slope**
  - The value of ***b*** is the **y-intercept**
  - ***r*** is the correlation coefficient
  - r^2 is the coefficient of determination

5. Interpreting and Visualizing

- Write down the equation of the line in slope intercept form.
- Press **Y=** and enter the equation under Y1. (Clear all other equations.)
- Press **GRAPH** and the line will be graphed through the data points.

Questions ???

Interpretation in Context

- Regression Equation:

$$y=1.5*x - 96.9$$

Water Consumption =

1.5*Temperature - 96.9

Interpretation in Context

- Slope = 1.5 (ounces)/(degrees F)
  - for each 1 degree F increase in temperature, you expect an increase of 1.5 ounces of water drank.

Interpretation in Context

y-intercept = -96.9

- For this example, when the temperature is 0 degrees F, then a person would drink about -97 ounces of water.
- That does not make any sense!
- Our model is not applicable for x=0.

Prediction Example

- **Predict** the amount of water a person would drink when the temperature is **95 degrees F.**

- **Solution:** Substitute the value of x=95 (degrees F) into the regression equation and solve for y (water consumption).

  If x=95, y=1.5*95 - 96.9 = **45.6 ounces.**

Strength of the Association: r^2

- Coefficient of Determination – r^2

- **General Interpretation:** The coefficient of determination tells the **percent of the variation** in the response variable that is **explained** (determined) by the model and the explanatory variable.

Interpretation of r^2

- Example: r^2 =92.7%.
- **Interpretation:**
  - Almost 93% of the variability in the amount of water consumed is explained by outside temperature using this model.
  - Note: Therefore 7% of the variation in the amount of water consumed is not explained by this model using temperature.

Questions ???

Simple Linear Regression Model

The model for
simple linear regression is

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

There are mathematical assumptions behind the concepts that
we are covering today.

Formulas

Prediction Equation: $\hat{Y} = mX + b$

$$slope = m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

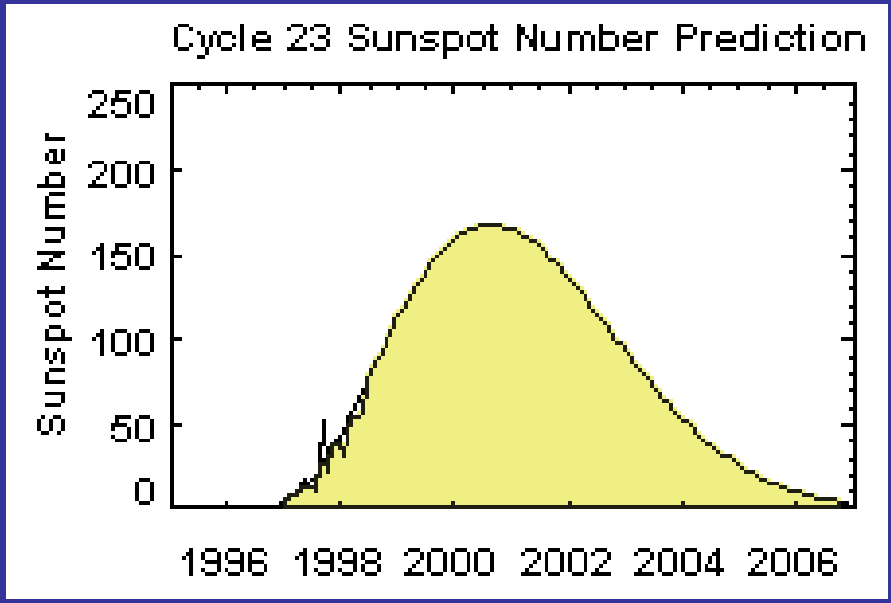
$$b = \bar{y} - m\bar{x}$$

Real Life Applications

Cost Estimating for Future Space Flight Vehicles (Multiple Regression)

Nonlinear Application

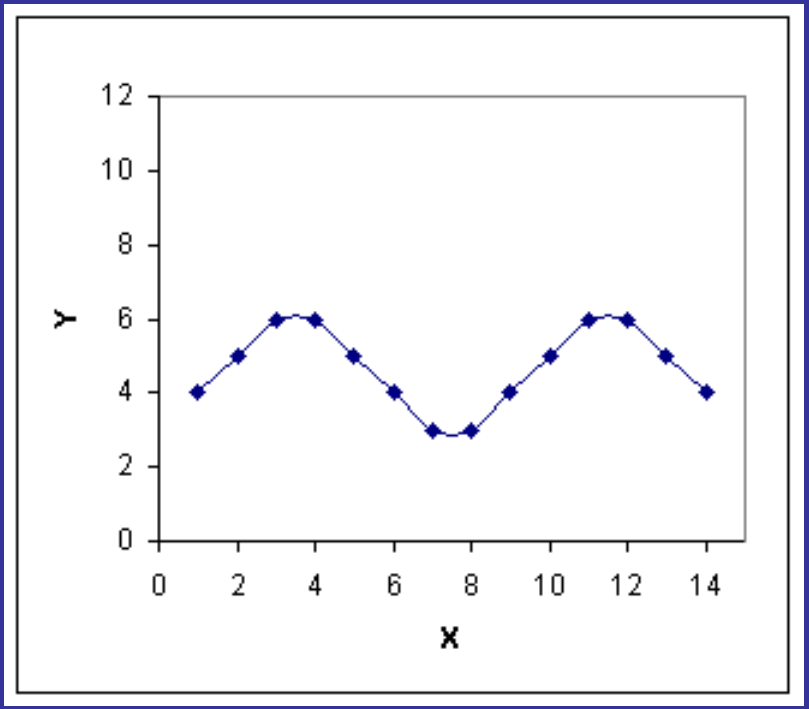
Predicting when Solar Maximum Will Occur



http://science.msfc.nasa.gov/ssl/pad/solar/predict.htm

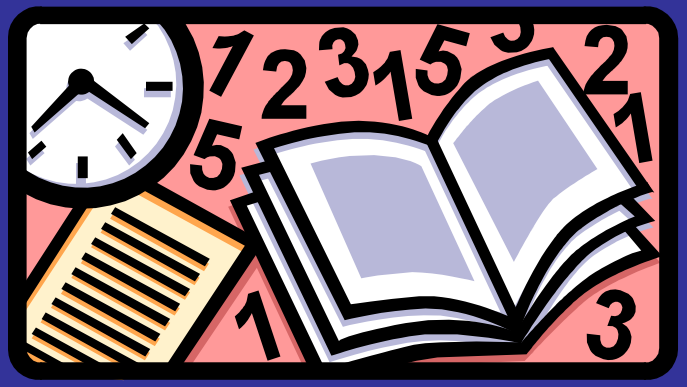
Real Life Applications

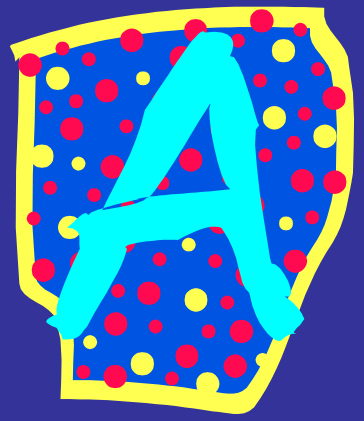
- Estimating Seasonal Sales for Department Stores (Periodic)



Real Life Applications

- Predicting Student Grades Based on Time Spent Studying





Real Life Applications

- . . .
- What ideas can you think of?
- What ideas can you think of that your students will relate to?

Practice Problems

- Measure Height vs. Arm Span
- Find line of best fit for height.
- Predict height for one student not in data set. Check predictability of model.

Practice Problems

- Is there any correlation between shoe size and height?

- Does gender make a difference in this analysis?

Practice Problems

- Can the number of points scored in a basketball game be predicted by
- The time a player plays in the game?
  - By the player's height?
  - Idea modified from Steven King, Aiken, SC. NCTM presentation 1997.)

Resources

- *Data Analysis and Statistics*. Curriculum and Evaluation Standards for School Mathematics. Addenda Series, Grades 9-12. NCTM. 1992.

- *Data and Story Library*. Internet Website. http://lib.stat.cmu.edu/DASL/ 2001.

Internet Resources

- Correlation
  - **Guessing Correlations** - An interactive site that allows you to try to match correlation coefficients to scatterplots. University of Illinois, Urbanna Champaign Statistics Program. http://www.stat.uiuc.edu/~stat100/java/guess/GCApplet.html

Internet Resources

- Regression
  - **Effects of adding an Outlier**.

    W. West, University of South Carolina.

    http://www.stat.sc.edu/~west/javahtml/Regression.html

Internet Resources

- Regression
  - **Estimate the Regression Line**. Compare the mean square error from different regression lines. Can you find the minimum mean square error? Rice University Virtual Statistics Lab. http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html

Internet Resources: Data Sets

- **Data and Story Library.**

  Excellent source for small data sets. Search for specific statistical methods (e.g. boxplots, regression) or for data concerning a specific field of interest (e.g. health, environment, sports). http://lib.stat.cmu.edu/DASL/

Internet Resources: Data Sets

- **FEDSTATS.** "The gateway to statistics from over 100 U.S. Federal agencies" http://www.fedstats.gov/

- **"Kid's Pages."** (not all related to statistics) http://www.fedstats.gov/kids.html

Internet Resources

- **Other**
  - Statistics Applets. Using Web Applets to Assist in Statistics Instruction. Robin Lock, St. Lawrence University. http://it.stlawu.edu/~rlock/maa99/

Internet Resources

- **Other**
  - Ten Websites Every Statistics Instructor Should Bookmark. Robin Lock, St. Lawrence University. http://it.stlawu.edu/~rlock/10sites.html

For More Information…

On-line version of this presentation
http://www.mtsu.edu/~stats/corregpres/index.html

More information about regression
Visit **STATS @ MTSU** web site
http://www.mtsu.edu/~stats