# CS130 Regression

Winter 2014

# Regression Analysis

- Regression analysis:
  - usually falls under statistics and mathematical **modeling**
  - is a form of statistical analysis used in **forecasting**
  - estimates the relationship between variables
    - **Allows predictions**
- During regression analysis, we need to fit functions to data.
  - What function best describes this data?

# Regression Analysis

- Trendlines are used to graphically display trends in data and to analyze problems of prediction.
- Draw a line that best fits the data.
- Regression analysis allows you to extend a trendline in a chart beyond the actual data to predict values
- Place the line such that the distance from each data point to the line is minimized.

# Regression Analysis

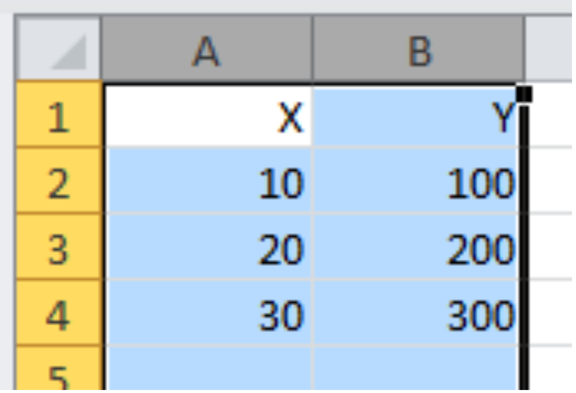
- There a many types of regression models, the most common is linear regression

- In linear regression, we try to find a straight line that best fits our data.
  - Plot data using Excel's XY or scatter chart.
  - Add the trendline to the chart

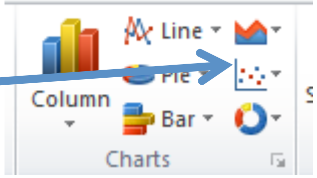
# Regression Analysis using Excel

Problem 7.1

Create the following workshe
Select both columns of data
Select the Insert tab

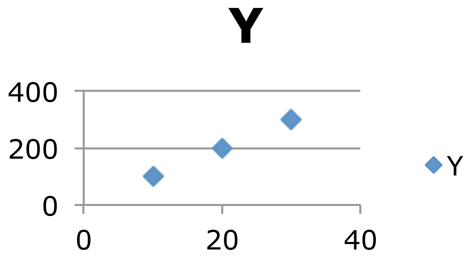


|   | A | B |
|---|---|---|
| 1 | X | Y |
| 2 | 10 | 100 |
| 3 | 20 | 200 |
| 4 | 30 | 300 |
| 5 |   |   |



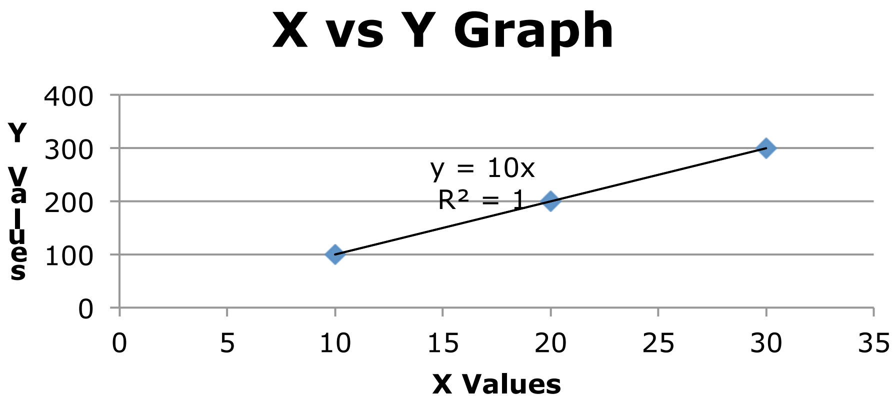
Select the ScatterPlot

Results



# Add Trendline & Equation

- Dress up the graph using the Layout tab
  - Select Axes Titles to label the x & y-axis
  - Select Analysis to add a trendline, equation, and **R^2** value



- Change the Y value from 200 to 150. What do you notice?

# Problem 7.2

In the CS130 Pub folder is a file called CandyBars.xls. Copy this file to your Desktop, open it and do the following.

1. Create a ScatterPlot of the data Carbohydrates and Sugars. Which goes on the X-Axis? Why?
2. Add a trendline to your chart, display the function or equation, and display the R^2 value
3. Is the function a good predictor? Why or Why not?
4. What is the amount of sugars (in grams) that we can expect from a candy bar with 60 grams of carbohydrates?
5. Add an empty column after name. In that column, place an asterisk for foods that have a carbohydrate count of 40 grams or higher and a sugar count of 35 grams or higher.
6. Turn on the **AutoFilter** and find out the number of M&M/Mars candy that fits these criteria.

# Nonlinear Regression

- Often times, relationships are nonlinear and we need a different type of graph to fit the data.
- Excel provides us with different types of nonlinear functions that we can use to fit data. These functions include:
  - Polynomial
  - Exponential
  - Logarithmic
  - Power

# Problem 7.3

## Copy FluSeason2012_2013
## from CS130 Public to your desktop

http://www.cdc.gov/flu/weekly/weeklyarchives2012-2013/data/whoAllregt11.htm

The flu season can be broken into two phase, flu growth and flu decline.

1. Fit different types of nonlinear functions to the growth data
2. Which works best?
3. How do we know?

| Flu Growth | | | Flu Decline | |
|---|---|---|---|---|
| Total Infections | Week | | Total Infections | Week |
| 163 | 1 | | 6425 | 14 |
| 197 | 2 | | 6832 | 15 |
| 300 | 3 | | 5892 | 16 |
| 339 | 4 | | 5093 | 17 |
| 409 | 5 | | 4029 | 18 |
| 619 | 6 | | 3056 | 19 |
| 851 | 7 | | 2376 | 20 |
| 1365 | 8 | | 2052 | 21 |
| 2030 | 9 | | 1731 | 22 |
| 3422 | 10 | | 1509 | 23 |
| 4561 | 11 | | 899 | 24 |
| 5891 | 12 | | | |
| 6442 | 13 | | | |

# Problem 7.3 Continued

1. If the growth phase did not end, how many infections would we expect in week 15?

2. If the growth phase did not end, in what week would we expect 10,000 infections?

# Solving Exponential and Logarithmic Equations

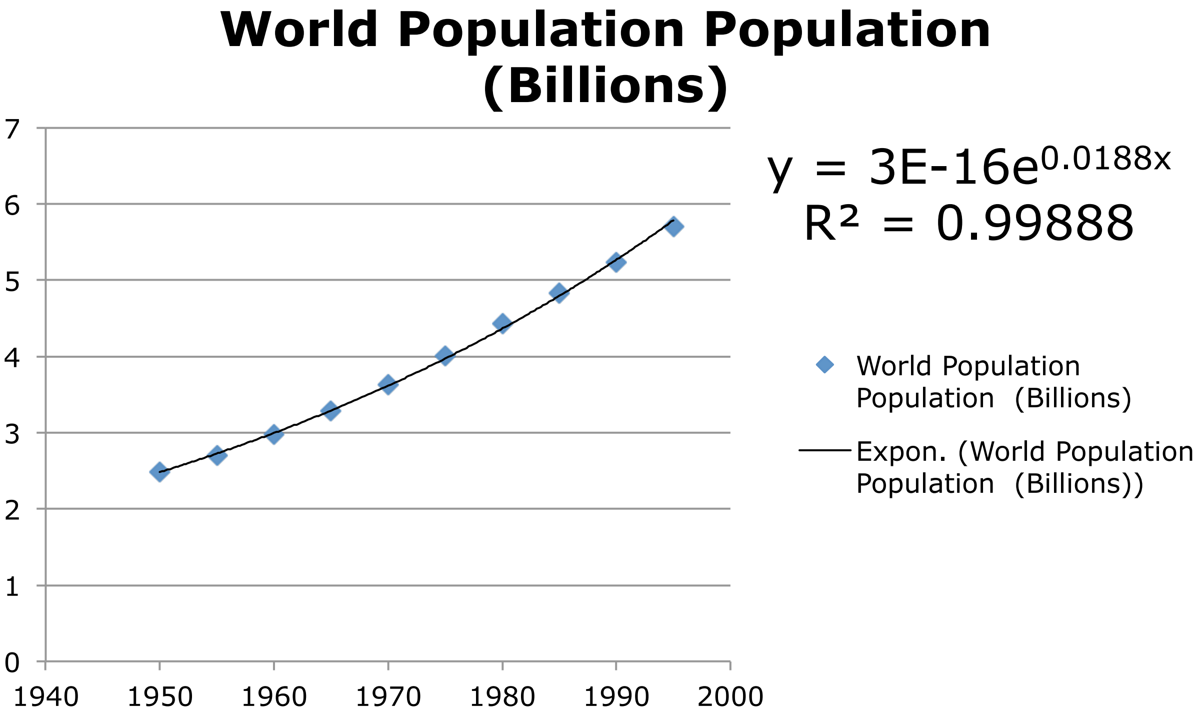
- Recall that to solve an equation of the form $y = ae^{bx}$ for x (where a and b are just constants), you first divide by a to obtain $y/a = e^{bx}$ . Now, you must take the natural logarithm of each side to obtain $\ln(y/a)=bx$. Dividing by b yields $x = (1/b)\ln(y/a)$.

- Recall that to solve an equation of the form $y = a \ln(bx)$ for x (where a and b are just constants), you again divide by a to obtain $y/a = \ln(bx)$. Now, you must exponentiate each side to obtain $e^{y/a} = bx$. Dividing by b yields $x = (1/b)e^{y/a}$ .

# Problem 7.4

http://zeus.cs.pacificu.edu/shereen/cs130w14/WorldPop.html

- Import this data into Excel and run an exponential regression.



The equation contains a good deal of rounding.

We know this from **E-16**

In order to use the equation to predict values:

Right Click Equation
Format Trendline Label
Number
Decimal Places: 18

# 7.4 Continued

- What is the predicted population in 2000?
- When will the population hit 7.0 billion people?
- Check WorldOMeters to see when the world hit 7 billion people.  How accurate was the model?

http://www.worldometers.info/world-population/

# Problem 7.5

The following data is from an actual study that considered how memory decreases with time.

- Read a list of 20 words slowly aloud
- later, at different time intervals, how many can you recognize?
- The percentage, P, of words recognized was recorded as a function of the time t elapsed in minutes.

# Problem 7.5 Continued

http://zeus.cs.pacificu.edu/shereen/cs130w14/Problem7.5.html

| T,min | 5 | 15 | 30 | 60 | 120 | 240 | 480 | 720 | 2880 | 5760 |
|---|---|---|---|---|---|---|---|---|---|---|
| P% | 73.0 | 61.7 | 58.3 | 55.7 | 50.3 | 46.7 | 38.3 | 29.0 | 24.0 | 18.7 |

1. What is the logarithmic trendline for the given data?
2. At what time T can we expect 40% of the words to be remembered? In order to solve this problem, rewrite the logarithmic equation solving for x. Then using Excel, find the answer to the given question.
3. Check your answer using Goal Seek. The two answers should be very close.