



CS130 Regression

Spring 2013

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 are 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

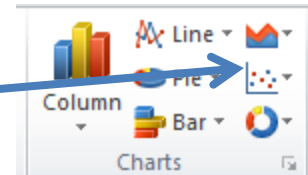
Regression Analysis using Excel

Problem 7.1

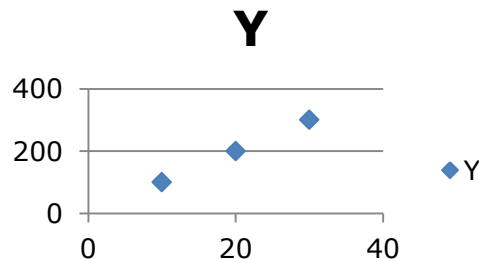
Create the following worksheet
Select both columns of data
Select the Insert tab

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

Select the ScatterPlot

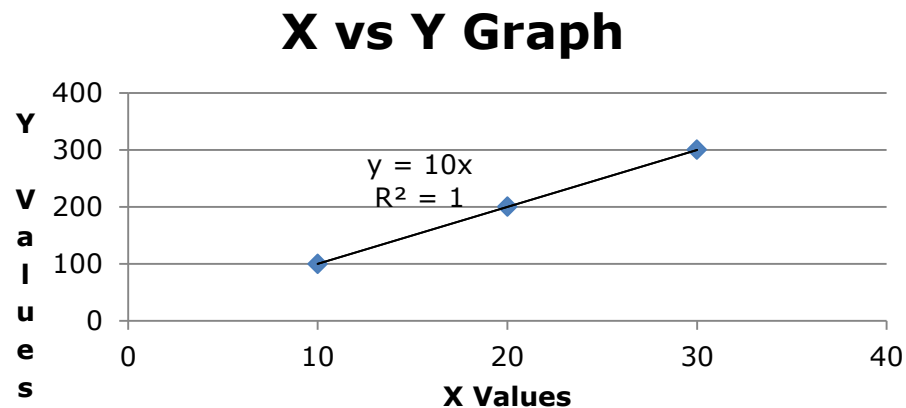


Results



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 AIDS.csv from CS130 Public to your desktop

Open with Excel

Save as AIDS.xlsx

Let us consider the following data which represents the number of deaths, N , from AIDS in the United States from 1981 to 1996, where t denotes the number of years after 1980.

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

t	N
1	159
2	622
3	2130
4	5635
5	12607
6	24717
7	41129
8	62248
9	90039
10	121577
11	158193
12	199287
13	243923
14	292586
15	340957
16	375904

Problem 7.3 Continued

1. What is the predicted number of deaths from AIDS in 1997?
 - Actual number of AIDS deaths in the US in 1997:
~18,000
2. In what year can we expect 1,000,000 deaths from AIDS?

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

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 7.4 Continued

zeus.cs.pacificu.edu/chadd/cs130s13/Problem74.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.