Regression Analysis

Regression Analysis

Regression analysis is a form of statistical analysis used for forecasting. Regression analysis estimates the relationship between variables, so that a particular variable can be predicted from one or more other variables. During regression analysis, we need to fit functions to data.

Trendlines are used to graphically display trends in data and to analyze problems of prediction. In other words we try to draw a line that best fits the data. By using regression analysis, you can extend a trendline in a chart beyond the actual data to predict future values.

This subject usually falls under statistics and mathematical modeling and can be applied to many different scientific and business applications. Understanding the various formulas for regression is beyond the scope of this class. However, you should understand that the line should be placed such that the distance or variation from each data point to the line is minimized.

Linear Regression

In linear regression we try to find a straight line that best fits our data. We first need to plot our data using Excel's XY or scatter chart. We then add the trendline to the chart and use the function to predict future values for our data.

The detailed steps are:

1. Enter the data in an Excel worksheet and select the data you want to plot.
2. Click on the chart wizard.
3. Choose XY (scatter) plot.
4. Check that the data range is correct.
5. Enter the titles and labels.
6. Click on the chart then select CHART from menu bar and ADD TRENDLINE from this menu.
7. From the menu that appears, select the type of function that you would like to use for your model. In this example we will use the default, which is LINEAR REGRESSION.
8. In order to have Excel display the equation of our regression line and the correlation coefficient, you need to click on the OPTIONS tab within the Add Trendline screen above, click on these two options, and then press OK.

You should be rewarded with a graph, equation and regression coefficient.

4.1 Exercise

In the CS130 Pub folder (in a folder called DataFiles) is a file called Candy Bars.xls. Copy this file to your folder, open it and do the following.

- **Part I:** Create a ScatterPlot of the data Carbohydrates and Sugars.

- **Part II:** Add a trendline to your chart and display the function or equation.

  Write the equation here: ___________________________

- **Part III:** What is the amount of sugars (in grams) that we can expect from a candy bar with 60 grams of carbohydrates?

  Write your answer here: ___________________________

- **Part IV:** 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.

- **Part V:** Turn on the AutoFilter and find out the number of M&M/Mars candy that fits these criteria.

Regression Coefficient

The regression coefficient, also known as the R-squared value, is an indicator that ranges in value from 0 to 1 and reveals how closely the estimated values for the trendline correspond to your actual data. A trendline is most reliable when its R-squared value is at or near 1.

Question: What is the regression coefficient in the exercise 4.1?
Answer: ________________________

Question: How well does the trendline represent the data in exercise 4.1?
Answer: ____

____

____

Nonlinear Regression

Often, 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 and power.

4.2 Exercise

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.

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

After typing the data into an Excel spreadsheet, try to fit different types of nonlinear functions to the data. Which works the best? How do we know?

Question: What is the predicted number of deaths from AIDS in 1997?
Answer: ____

____

Question: In what year can we expect 1,000,000 deaths from aids?
Answer: ____

____

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 $x = (1/b)e^{y/a}$.

4.3 Exercise

The following data is from an actual study that considered how memory decreases with time. The subjects each read a list of 20 words slowly aloud, and later, at different time intervals, were shown a list of 40 words containing the 20 words that he or she had read. The percentage, P, of words recognized was recorded as a function of the time t elapsed in minutes. The table below shows the averages for 5 different subjects.

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

Question: What is the logarithmic trendline equation for the given data?

Answer: ____

Question: 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.

Answer: ____

Problem: Check your answer using Goal Seek. The two answers should be very close.