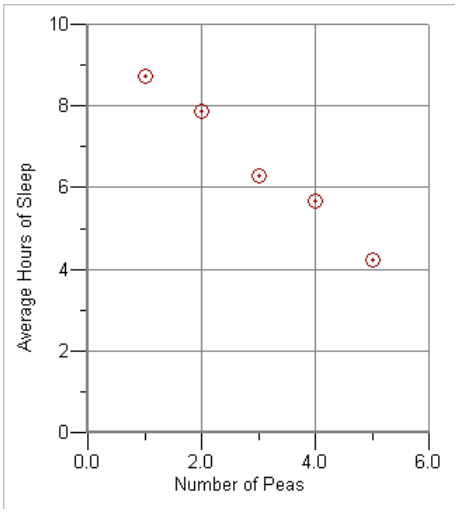
# The Mathematical Modeling of Experimental Data

In an experiment we usually measure a response for one or more variables as a function of a variable whose value is directly under our control. In the language of experimental design, the variable under our control is the independent variable and the variables whose values we measure are dependent variables. For example, consider the following hypothetical data for an experiment designed to determine the effect on a Princess's sleep of placing peas under her mattress.

| Number of Peas Under Mattress | Average Hours of Sleep Obtained by Princess |
|---|---|
| 1 | 8.72 |
| 2 | 7.86 |
| 3 | 6.29 |
| 4 | 5.68 |
| 5 | 4.22 |

In this experiment the number of peas under the mattress is the independent variable; that is, this is the variable under our control. The average hours of sleep, therefore, is the dependent variable. A graph of this data (see below to the right) shows what appears to be an inverse linear relationship between then number of peas placed under the mattress and the average hours of sleep. Seeing this relationship we might ask questions such as "What is the relationship between the average hours of sleep and the number of peas placed under a mattress?" or "If we place seven peas under the mattress, how many hours might the Princess sleep?"




**Regression Analysis**. To answer questions such as those suggested above requires a suitable mathematical equation that appropriately models the data. This is the realm of a regression analysis. For a straight-line relationship the model equation is

$$Y = \beta_0 + \beta_1 X$$

where $Y$ (the dependent variable) is the average hours of sleep, $X$ (the independent variable) is the number of peas placed under the mattress, $\beta_0$ is the average hours of sleep in the absence of any peas (the value of $Y$ when $X$ is zero, or the $y$-intercept) and $\beta_1$ is the average hours of sleep lost per pea (which also is the slope of the line or the rate of change of $Y$ relative to $X$; that is, $\Delta Y/\Delta X$). The terms $\beta_0$ and $\beta_1$ are consider adjustable fitting parameters of the model. The goal of a regression analysis is to find the best values for $\beta_0$ and $\beta_1$ such that the net difference between the experimental values of $Y$ and those values predicted by the model is as small as possible.[1] The mathematical details of how this is accomplished are too involved for this course. Fortunately, we have access to software packages that can carry out the analysis for us.

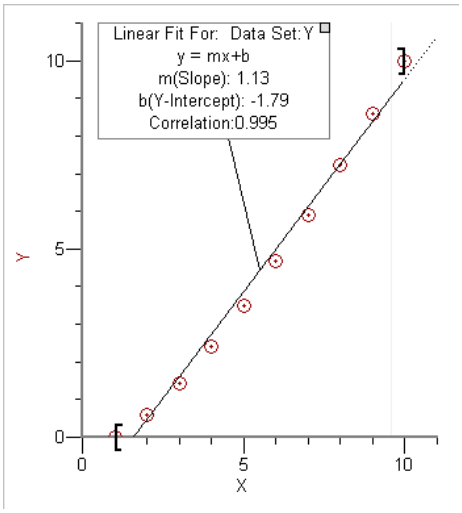
Of course you can fit a straight-line to any set of data, even if the data clearly are not linear. For this reason it is important to examine the results of a regression analysis and determine whether your model is reasonable. One common way to evaluate what is often called the model's "goodness of fit"

---

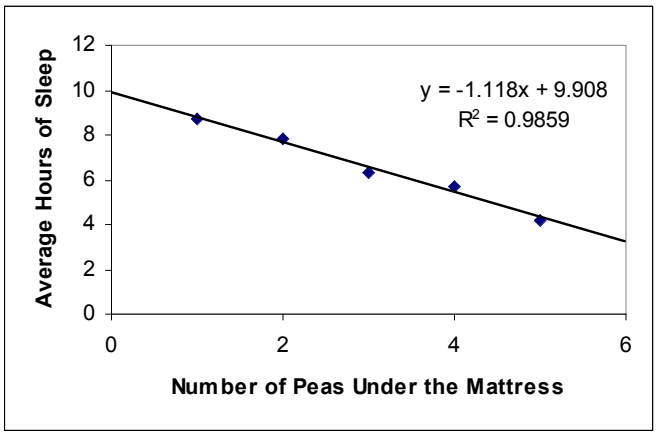
[1] You may be more familiar with expressing a straight-line in the form $Y = mX + b$, where $m$ is the slope and $b$ is the $y$-intercept. The use of $\beta_0$ and $\beta_1$, however, is the more standard statistical form of the equation.

is to look at the correlation coefficient, $R$, or the coefficient of determination, $R^2$, which are two measures of the degree to which the model explains the data. A value of $R$ close to +1 or to –1 (or an $R^2$ close to +1) suggests the model does a good job of explaining the data and a value for $R$ or for $R^2$ close to 0 suggests that the model is inappropriate.

Unfortunately, the correlation coefficient and the coefficient of determination are not always a very sensitive measure of a model's suitability. Large values for $R$ or for $R^2$ can falsely lead you to assume that a model provides an accurate description of the data. A much better choice is to graph both the data and the predicted model and examine them critically. If a model is appropriate then the model should fit closely the data with individual data points randomly scattered around the model's predicted curve. Note that although the example on the left has a very favorable value for $R^2$, the data clearly show evidence of curvature with values of $Y$ for low and for high values of $X$ found above the model's curve and values of $Y$ for intermediate values of X falling below the model's curve. A quadratic model of the form $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ is probably a better choice for this data.



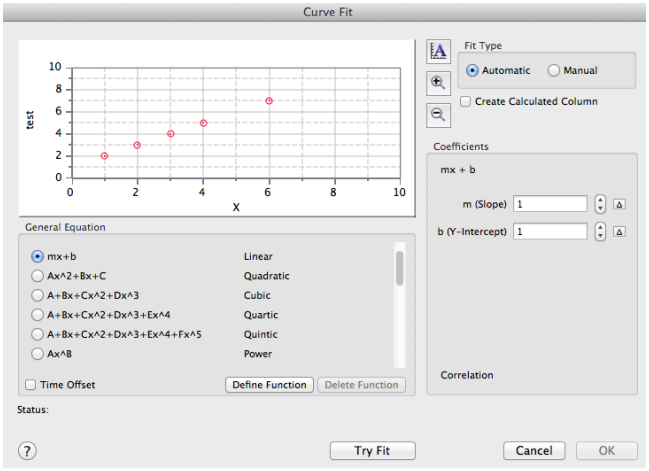

**Using Excel for a Regression Analysis**. You may already be familiar with using Excel to complete a regression analysis and know how to add the resulting model's curve to a chart. If not, then here are a few instructions. Begin by creating your chart using the data provided above. Click on the chart's data and select Chart:Add Trendline. Note that there are six options, one of which is for a linear trend, which is the one you wish to use here. Select the options tab and click on the appropriate boxes to add the equation to your chart and to display the $R^2$ value. By default, Excel only displays the regression line from the first value to the last value on the $x$-axis. If you wish to extend the regression line in either direction, you can do so using the forecast section of this window by indicating how many units on the $x$-axis to extend the line. When complete your chart should look similar to that shown below. Note that this example extends the model one unit in either direction along the $x$-axis; that is to values of $X$ corresponding to zero and six peas.

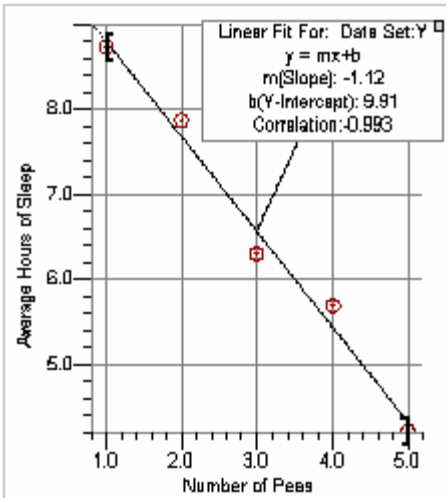



Using Excel to complete a regression analysis is not without its limitations. Excel has a limited choice of mathematical models and does not allow you to specify other models. Furthermore, in some cases the equation provided by Excel includes too few significant figures for $\beta_0$ and $\beta_1$ to be of practical use. To increase the number of significant figures, click on the equation and select Format:Selected Trendline Labels, chose the option for Number and specify the appropriate format.

**Using Logger Pro for a Regression Analysis**. LoggerPro is a better choice for completing a regression analysis. Here are some instructions. Begin by creating a graph of the data provided above. Select Analyze:Curve Fit from the main

menu, providing the Curve Fit window shown to the right. The scrolling menu at the bottom of the window provides a range of functions. Select the appropriate function and click on the Try Fit button to preview the result. If acceptable, click on OK to accept the result. You should obtain a result similar to that shown below. Note that for a straight-line model Graphical Analysis provide the correlation coefficient ($R$) instead of the coefficient of determination ($R^2$). For other models, Graphical Analysis reports the root-mean-square-error (RMSE), which is defined as



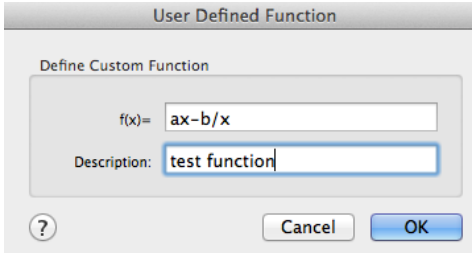




$$\text{RMSE} = \sqrt{\frac{\Sigma(y_i - \hat{y}_i)}{n}}$$

where $y_i$ is the experimentally determined value the $i^{\text{th}}$ value of $x$, $\hat{y}_i$, which is read as "y-hat," is the model's predicted value for $y_i$, and $n$ is the number of measurements. The smaller the root-mean-square-error, the better the fit between the model and the data.

LoggerPro provides two important features that are not available with Excel's trendline function. First, clicking and dragging the brackets on the graph (the **[** and **]** in the figure to the left) changes the portion of your data included in the regression analysis; the model's predicted equation adjusts automatically. Second, you can use the Define Function button on the Curve Fit window (see figure at the top of the page) to use a model that is not included with the software. Shown to the right, for example, is the format for entering the model

$$Y = aX - b/X$$




where $a$ and $b$ are the coefficients used to fit the model to the data.

**Interpolating and Extrapolating From a Model**. The reason for developing a regression model is to predict the value of the independent variable (or the dependent variable) for samples where its value is unknown. For example, we have shown (se figure above) that a Princess's average hours of sleep is a function of the number of peas we place under her mattress.

$$\text{Avg. Hours Sleep} = -1.12 \times \text{Number of peas} + 9.91$$

We can use this model in two ways. If we know how many peas we plan to place under her mattress, we can predict how long the princess will sleep. Alternatively, if we measure the number of hours the Princess sleeps on a given night, we can predict how many peas were under her mattress. These are powerful and useful applications of a model; however, when using a model to make predictions

we need to be careful when interpreting the results. Here we need to make an important distinction between interpolation and extrapolation.

In developing our model we used samples of 1, 2, 3, 4, and 5 peas. Based on our analysis of this data, we have every confidence that the mathematical model works well for this range of peas and hours of sleep. If we limit the model to making a prediction within this range, a process called interpolation, our confidence in the prediction's accuracy is high. For example, if we determine that the Princess slept 6.0 hours last night, then we can predict that there were 3.5 peas under her mattress and be confident in this prediction.

Extending the model to values of the dependent variable and the independent variable that we did not study, a process called extrapolation, is possible but more susceptible to uncertainty. If the Princess sleeps 10 hours our mathematical model suggests there probably were no peas placed under her mattress. This extrapolation of our model to smaller values of the independent variable seems reasonable as there is no reason to believe that the linear behavior between 1 and 5 peas does not hold between 0 and 1 peas.

Can we safely extrapolate the model to larger values of the dependent variable or independent variable? What is our prediction, for example, if we place 10 peas under the Princess's mattress? Using our model, we predict that the Princess will sleep for –1.29 hours, a result that is impossible. We clearly cannot extrapolate our model this far. Given this contradiction, it is tempting to modify our model by assuming that it is valid until the dependent variable reaches zero. Such an assumption, however, is still fraught with potential uncertainty. Suppose the Princess sleeps for 2.0 hours. Extrapolating our model leads us to predict that there are 7 peas under the mattress; however, it also is possible that the Princess will sleep a minimum of two hours regardless of the number of peas under the mattress. If true, then we cannot extrapolate the model to 2.0 hours or less of sleep.

When building models for a system it is important to consider how you plan to use the model and, if possible and practicable, to ensure that the range of values for the independent variable spans the range of values you wish to model. In this way your predictions rely on interpolations and not extrapolations. If an extrapolation is necessary, be sure to consider its limitations.