Mathematical Modeling of Data

In an experiment we usually measure a response for one variable as a function of a second 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 variable whose value we measure is the *dependent variable*. For example, consider the following hypothetical data for an experiment designed to determine the effect on a Princess' sleep (the dependent variable) of placing peas (the independent variable) under her mattress.

number of peas	average hours of sleep
1	8.72
2	7.86
3	6.29
4	5.68
5	4.22

A quick scan of this data suggests there is an inverse linear relationship between the number of peas and the average hours of sleep: fewer peas results in more sleep. Seeing this relationship we might ask questions such as "What is the mathematical relationship between the average hours of sleep and the number of peas placed under the 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 mathematical equation that models the data. This is the realm of a regression analysis. For a straight-line relationship the model equation is

$$y = \beta_0 + \beta_1 x \tag{1}$$

where y (the dependent variable) is the average hours of sleep, x (the independent variable) is the number of peas placed under the mattress, β_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 β_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, $\frac{\Delta y}{\Delta x}$). The terms β_0 and β_1 are adjustable fitting parameters of the model. The goal of a regression analysis is to find the best values for β_0 and for β_1 such that the net difference between the experimental values of y and those values predicted by the model is as small as possible. The mathematical details of how this is accomplished are too involved for this course; nevertheless, both Excel and LoggerPro have functions that will complete a regression analysis and add the model to a plot of your data, as shown in Figure 1, where the mathematical model is

$$\text{average hours sleep} = -1.12 \times \text{ number of peas} + 9.91 \tag{2}$$

Is My Regression Model A Good Model?

Of course we 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 the model is reasonable. One common method to evaluate what often is called the model's "goodness of fit" is to look at the correlation coefficient, R, or the coefficient of determination, R^2. 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 its data and a value for R or for R^2 close to 0

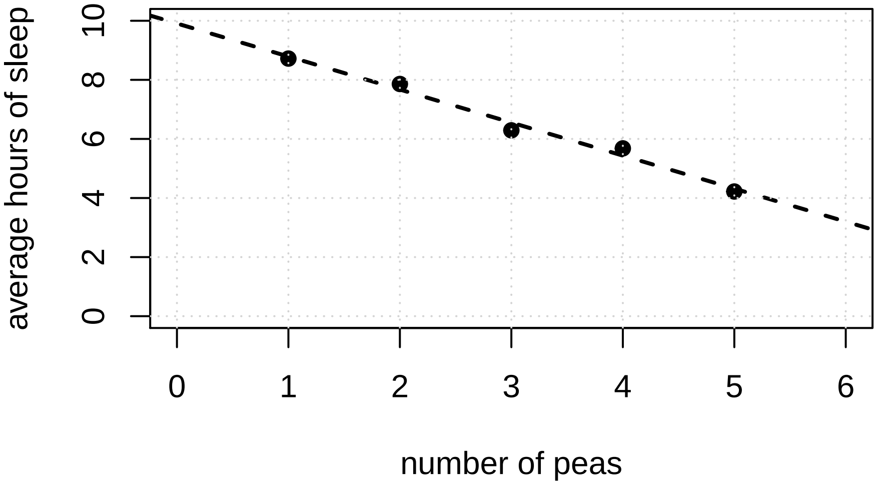

Figure 1: Average hours of sleep of the Princess as a function of the number of peas placed under her mattress. Each point is the average of three nights. The dashed line is equation 2 is the result of fitting equation 1 to the data.

suggests that the model is inappropriate. For the data in Figure 1, the values of R and R^2 are 0.986 and 0.981, respectively.

Another common measure of a model's goodness of fit is the root-mean-square error, which is defined as

$$\text{RMSE} = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n}}$$

where y_i is the experimental value for the i^{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, which essentially is the average difference between the data and the model—the better the fit between the model and the data. For the data in Figure 1, the RMSE is 0.080.

By themselves, the correlation coefficient, the coefficient of determination, and the root-mean-square error are not always useful measures of a model's suitability. Large values for R or for R^2, or a small RMSE may falsely lead you to assume that a model provides an accurate description of the data. Before you accept the result of any mathematical model, prepare a plot of the data and the predicted model and examine them critically. If the model is a good model then the regression line should closely fit the data with individual data points scattered randomly around the model's predicted curve. Note that although the regression model in Figure 2 has a favorable value for R^2, the data clearly show evidence of curvature with values of y for low and for high values of x falling below the model's curve and values of y for intermediate values of x falling above the model's curve. A quadratic model of the form $y = \beta_0 + \beta_1 x + \beta_0 x^2$ might be a better choice for this data.

Interpolating and Extrapolating From a Model

The reason for developing a regression model is to predict the value of the independent variable (or to predict the value of dependent variable) for a sample where its value is unknown. Recall that our model for the data in Figure 1 is

$$\text{average hours sleep} = -1.12 \times \text{ number of peas} + 9.91$$

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 placed under her mattress. These are powerful

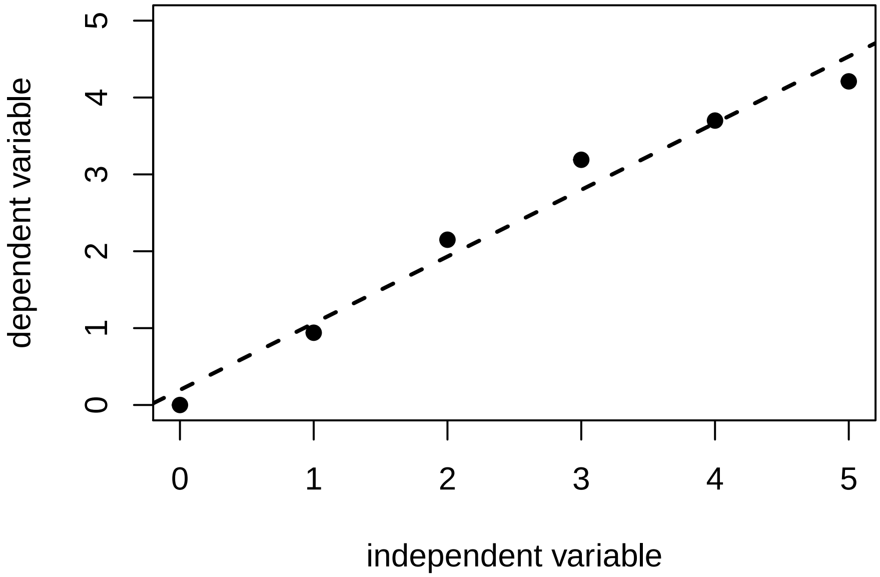

Figure 2: Example of data where a straight-line regression model provides a poor fit to the data even though the values for R and for R^2 are, respectively, 0.973 and 0.967.

and useful applications of a model; however, when when we use a model to a make prediction we need to be careful how we interpret the result. Here we need to make an important distinction between interpolation and extrapolation.

To develop the model in Figure 1 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, then our confidence in our 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 our model to values of the dependent variable and the independent variable that we did not study, a process called extrapolation, is possible but it is 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 the 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 you build a mathematical model it is important to consider how you plan to use the model and, if it is 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.