Introduction to Linear Regression

Prerequisites
Measures of Variability, Describing Bivariate Data

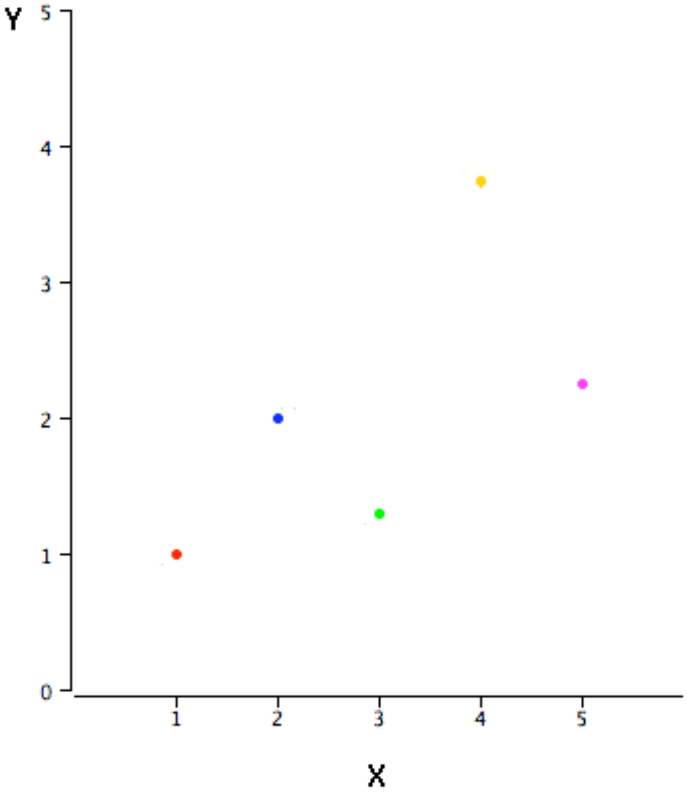
In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the *criterion variable* and is referred to as Y. The variable we are basing our predictions on is called the *predictor variable* and is referred to as X. When there is only one predictor variable, the prediction method is called *simple regression*. In simple linear regression, the topic of this section, the predictions of Y when plotted as a function of X form a straight line.

The example data in Table 1 are plotted in Figure 1. You can see that there is a positive relationship between X and Y. If you were going to predict Y from X, the higher the value of X, the higher your prediction of Y.

Table 1. Example data.

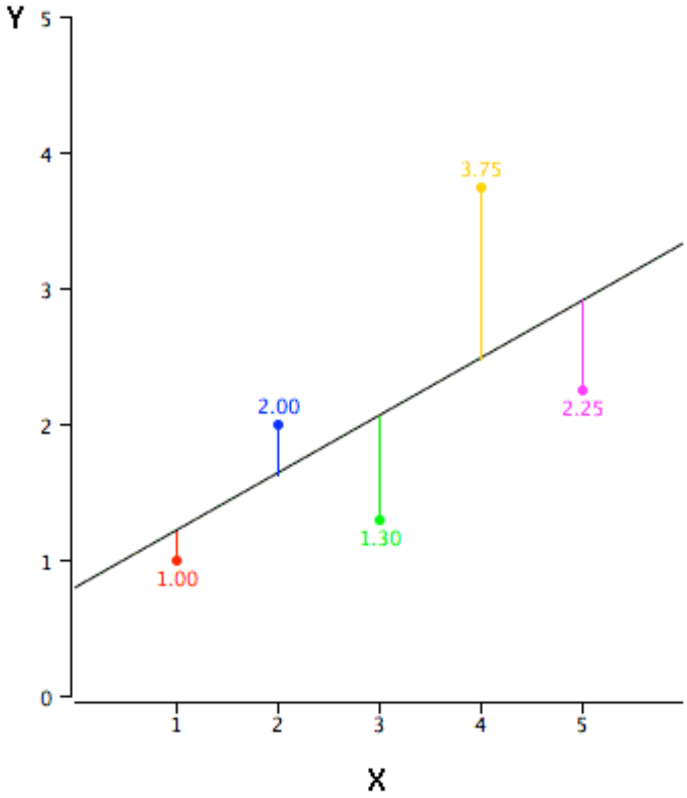
X	Y
1.00	1.00
2.00	2.00
3.00	1.30
4.00	3.75
5.00	2.25

Figure 1. A scatterplot of the example data.



Linear regression consists of finding the best-fitting straight line through the points. The best-fitting line is called a *regression line*. The black diagonal line in Figure 2 is the regression line and consists of the predicted score on Y for each possible value of X. The vertical lines from the points to the regression line represent the errors of prediction. As you can see, the red point is very near the regression line; its error of prediction is small. By contrast, the yellow point is much higher than the regression line and therefore its error of prediction is large.

Figure 2. A scatterplot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.



The error of prediction for a point is the value of the point minus the predicted value (the value on the line). Table 2 shows the predicted values (Y') and the errors of prediction (Y-Y'). For example, the first point has a Y of 1.00 and a predicted Y of 1.21. Therefore its error of prediction is -0.21.

Table 2. Example data.

X	Y	Y'	Y-Y'	$(Y-Y')^2$
1.00	1.00	1.210	-0.210	0.044
2.00	2.00	1.635	0.365	0.133
3.00	1.30	2.060	-0.760	0.578
4.00	3.75	2.485	1.265	1.600
5.00	2.25	2.910	-0.660	0.436

You may have noticed that we did not specify what is meant by "best fitting line." By far the most commonly used criterion for the best fitting line is the line that minimizes the sum of the squared errors of prediction. That is the criterion that was used to find the line in Figure 2. The last column in Table 2 shows the squared errors of prediction. The sum of the squared errors of prediction shown in Table 2 is lower than it would be for any other regression line.

The formula for a regression line is

$$Y' = bX + A$$

where Y' is the predicted score, b is the slope of the line, and A is the Y intercept. The equation for the line in Figure 2 is

$$Y' = 0.425X + 0.785$$

For X = 1,

$$Y' = (0.425)(1) + 0.785 = 1.21.$$

For X = 2,

$$Y' = (0.425)(2) + 0.785 = 1.64.$$

Computing the Regression Line

In the age of computers, the regression line is typically computed with statistical software. However, the calculations are relatively easy are given here for anyone who is interested. The calculations are based on the statistics shown in Table 3. MX is the mean of X, MY is the mean of Y, s_X is the standard deviation of X, s_Y is the *standard deviation* of Y, and r is the *correlation* between X and Y.

Formula for standard deviation
Formula for correlation

Table 1. Statistics for computing regression line

M_X	M_Y	s_X	s_Y	r
3	2.06	1.581	1.072	0.627

The slope (b) can be calculated as follows:

$$b = r\ sY/sX$$

and the intercept (A) can be calculated as

$$M_Y - bM_X.$$

For these data,

```
b = (0.627)(1.072)/1.581 = 0.425

A = 2.06 - (0.425)(3)=0.785
```

Note that the calculations have all been shown in terms of sample statistics rather than population parameters. The formulas are the same; simply use the parameter values for means, standard deviations, and the correlation.

A Real Example

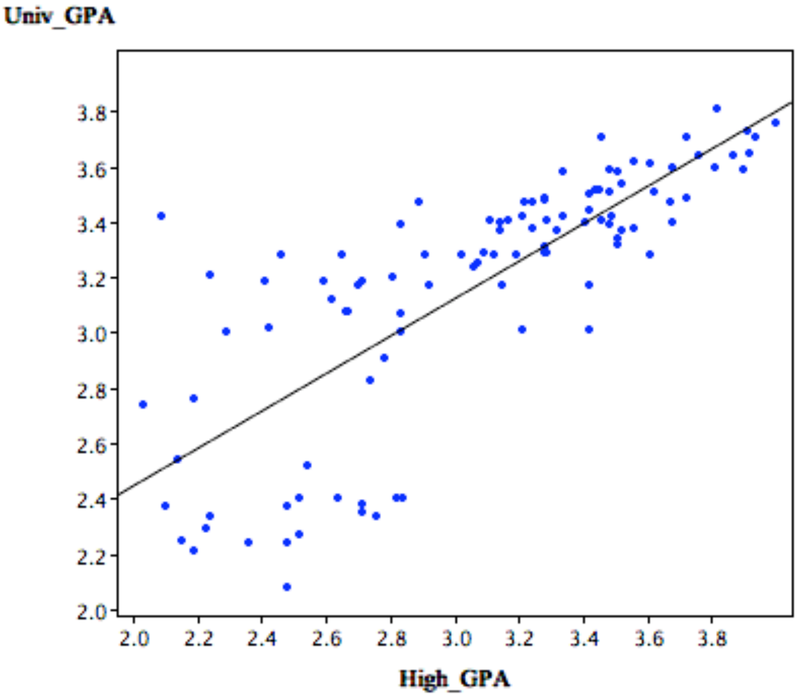
The case study, Predicting GPA contains high school and university grades for 105 computer science majors at a local state school. We now consider how we could predict a student's university GPA if we knew his or her high school GPA. Figure 3 shows a scatterplot of University GPA as a function of high school GPA. You can see from the figure that there is a strong positive relationship. The correlation is 0.78. The regression equation is

```
University GPA' = (0.675)(High School GPA) + 1.097
```

Therefore a student with a high school GPA of 3 would be predicted to have a university GPA of

```
University GPA' = (0.675)(3) + 1.097 = 3.12.
```

Figure 3. University GPA as a function of High School GPA.



Assumptions

It may surprise you, but the calculations shown in this section are assumption free. Of course, if the relationship between X and Y is not linear, a different shaped function could fit the data better. *Inferential statistics* in regression are based on several assumptions, and these assumptions are in a section of this chapter.