**Model-driven DSS cont.**

# 

# **1. Linear Regression**

# **2. Logistic Regression**

# **3. Ridge Regression**

# **4. Lasso Regression**

# **5. Polynomial Regression**

# **6. Bayesian Linear Regression**

# **Simple Linear Regression**

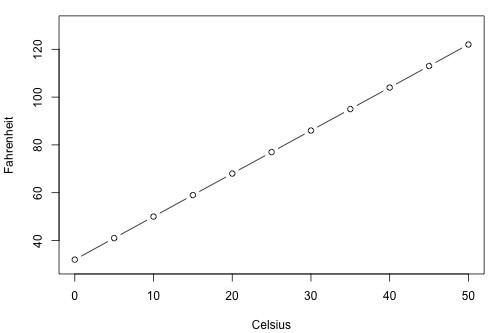
Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables.

One variable, denoted x, is regarded as the predictor, explanatory, or independent variable.

The other variable, denoted y, is regarded as the response, outcome, or dependent variable.

### **Types of relationships**

Before proceeding, we must clarify what types of relationships we won't study in this course, namely, **deterministic** (or **functional**) **relationships**. Here is an example of a deterministic relationship.



Note that the observed (*x*, *y*) data points fall directly on a line. As you may remember, the relationship between degrees Fahrenheit and degrees Celsius is known to be:

Fahr =95Cels+32

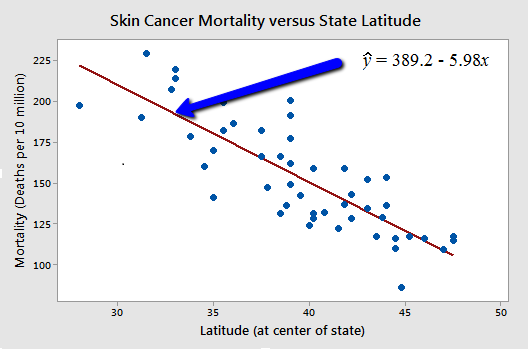
That is, if you know the temperature in degrees Celsius, you can use this equation to determine the temperature in degrees Fahrenheit *exactly*.

Here are some examples of other deterministic relationships that students from previous semesters have shared:

* Circumference = π × diameter
* Hooke's Law: *Y* = α + *βX*, where *Y* = amount of stretch in a spring, and *X* = applied weight.
* Ohm's Law: *I* = *V*/*r*, where *V* = voltage applied, *r* = resistance, and *I* = current.
* Boyle's Law: For a constant temperature, *P* = α/*V*, where *P* = pressure, α = constant for each gas, and *V* = volume of gas.

For each of these deterministic relationships, the equation *exactly* describes the relationship between the two variables. This course does not examine deterministic relationships. Instead, we are interested in **statistical relationships**, in which the relationship between the variables is not perfect.

Here is an example of a statistical relationship. The response variable *y* is the mortality due to skin cancer (number of deaths per 10 million people) and the predictor variable *x* is the latitude (degrees North) at the center of each of 49 states in the U.S. ([skincancer.txt](https://onlinecourses.science.psu.edu/stat501/sites/onlinecourses.science.psu.edu.stat501/files/data/skincancer.txt)) (The data were compiled in the 1950s, so Alaska and Hawaii were not yet states. And, Washington, D.C. is included in the data set even though it is not technically a state.)



Some other examples of statistical relationships might include:

* Height and weight — as height increases, you'd expect weight to increase, but not perfectly.
* Alcohol consumed and blood alcohol content — as alcohol consumption increases, you'd expect one's blood alcohol content to increase, but not perfectly.
* Vital lung capacity and pack-years of smoking — as amount of smoking increases (as quantified by the number of pack-years of smoking), you'd expect lung function (as quantified by vital lung capacity) to decrease, but not perfectly.
* Driving speed and gas mileage — as driving speed increases, you'd expect gas mileage to decrease, but not perfectly.

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.

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 |

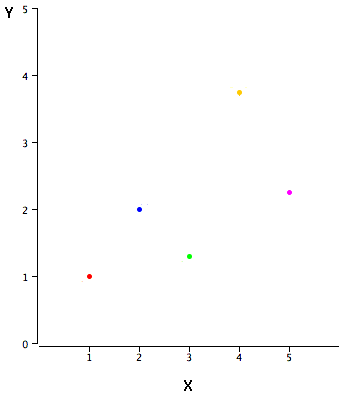


Figure 1. A scatter plot of the example data.

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.

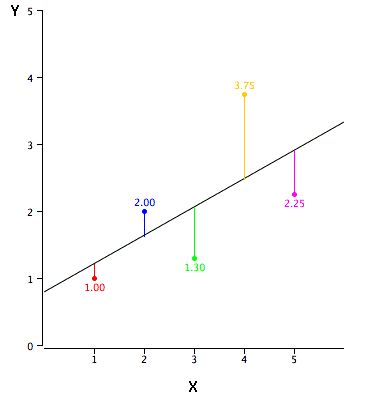


Figure 2. A scatter plot of the example data. The black line consists of the predictions, the points are the actual data, and the vertical lines between the points and the black line represent errors of prediction.

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 (called 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.

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, and 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, sX is the standard deviation of X, sY is the *standard deviation* of Y, and r is the *correlation* between X and Y.

Table 3. Statistics for computing the regression line.

| **MX** | **MY** | **sX** | **sY** | **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

A = MY - bMX.

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.

**Standardized Variables**

The regression equation is simpler if variables are *standardized* so that their means are equal to 0 and standard deviations are equal to 1, for then b = r and A = 0. This makes the regression line:

ZY' = (r)(ZX)

where ZY' is the predicted standard score for Y, r is the correlation, and ZX is the standardized score for X. Note that the slope of the regression equation for standardized variables is r.

**A Real Example**

The case study "[SAT and College GPA"](http://onlinestatbook.com/2/case_studies/sat.html) 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 scatter plot 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.

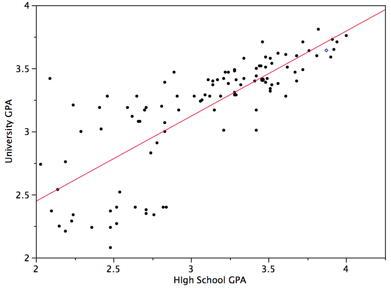
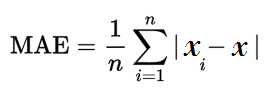


Figure 3. University GPA as a function of High School GPA.

## Mean Absolute Error

The **Mean Absolute Error**(MAE) is the average of all absolute errors. The formula is:

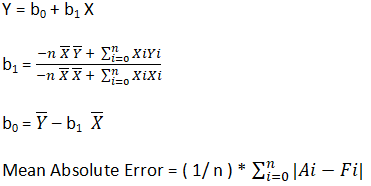


**Where**:

* n = the number of errors,
* Σ = summation symbol (which means “add them all up”),
* |xi – x| = the absolute errors.

The formula may look a little daunting, but the steps are easy:

1. Find all of your absolute errors, xi – x.
2. Add them all up.
3. Divide by the number of errors. For example, if you had 10 measurements, divide by 10.



**Example:**

Consider a dataset consisting of 10 instances and 2 attributes as income and price.  
a-Find the equation of the simple linear regression for this dataset.   
b-Calculate the mean absolute error.  
  
  
Income(TL)X Price (TL)Y  
 76 24  
 92 24  
 107 32  
 112 32  
 119 31  
 129 34  
 143 35  
 160 39  
 180 40  
 193 46  
  
where n is the total number of instances, A is the actual Y value and F is the forecasted Y value.