Virginia Cooperative Extension
Virginia Tech • Virginia State University

EXPERT REVIEWED

CNRE-184P

Prediction Equations: Intuition and Implementation in Forestry

Authored by Corey Green, Assistant Professor, Forest Resources and Environmental Conservation, Virginia Tech

Consider the following scenario: A forester is tasked with determining the value of the standing timber in a 75-acre, 25-year-old planted loblolly pine stand. Depending on the region, timber may be sold by the cubic foot, green or dry ton, board foot (the equivalent volume of a 1-inch-thick piece of wood, 12 inches square), or even by carbon content. The forester has simple tools for measuring the diameter at breast height (diameter at 4.5 feet above the ground, commonly referred to as DBH) and stem height (total height from ground to tip, or THT). How would the forester go about this? Cutting the trees down is not an option, so some way of predicting the standing content is needed.

Another scenario: A forester needs an estimate of the average height of trees in a given forest. Measuring height is not difficult, but it is more time consuming than measuring DBH. The forester decides to measure DBH on all trees but heights on just a subset. The forester now needs a way to determine the heights of the trees not measured.

Both of the scenarios above represent situations where something relatively difficult to measure is needed (e.g., tree volume) and only simple, easy measurements are available (e.g., DBH).

The objective of this article is to describe and demonstrate a powerful statistical tool called linear regression. This technique can be used to predict tree characteristics that are difficult to measure but important for forest management. Yellow poplar (*Liriodendron tulipifera*) data from Burkhart, Avery, and Bullock (2019) will be used to illustrate this statistical tool.

In figure 1 below, a scatterplot is used to show the relationship between two characteristics. Each open circle represents an x-y pair. The easier-to-measure characteristics for an individual are located on the x-axis (horizontal). The y-axis (vertical) represents the harder-to-measure characteristics for those same individuals. The red arrows refer to a single individual's x and y values.

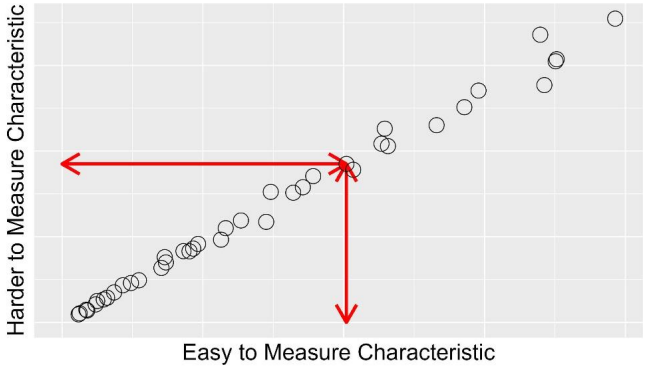

Figure 1. The relationship between two characteristics for individuals.

Looking at this relationship, it should be apparent that the two characteristics are strongly related. If you know x (the easier thing to measure), you should be able to predict y (the harder thing to measure). Unfortunately, this isn't always the case. Compare the two scatterplots in figure 2.

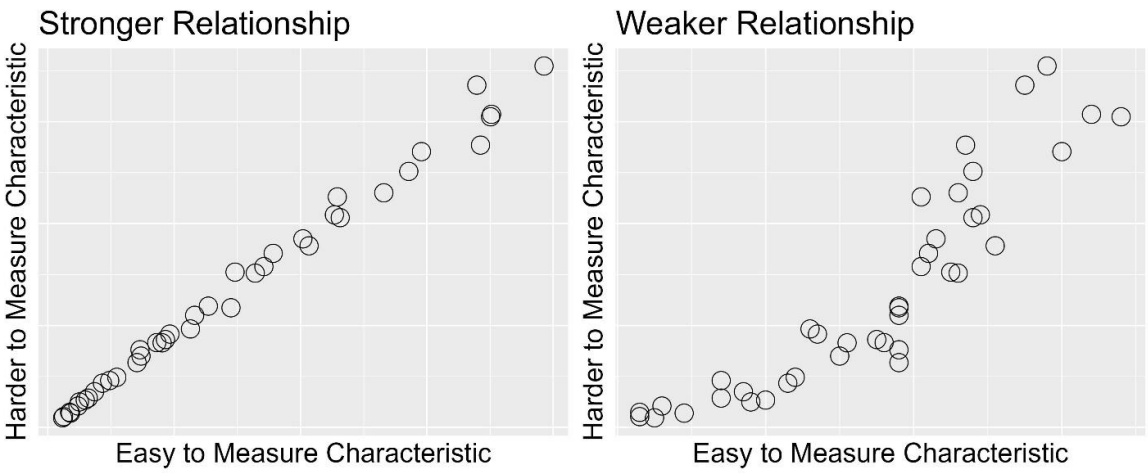

Figure 2. A comparison of two relationships.

Clearly, the left panel of figure 2 is a stronger relationship. You would be less confident using the relationship in the right panel to predict y from x.

How can we use these relationships? Return to the first scenario: You need to know the total cubic foot volume in a standing tree. One option would be to fell the tree and cut it into lots of small pieces. Each piece could then be submerged in water, and the displacement could be measured. This would result in a highly accurate measure of volume; however, this is highly impractical, especially if the tree needs to remain standing.

An alternative approach commonly used by foresters is to use an equation that can predict volume. After going through the challenging work of felling lots of trees and actually measuring the volume, simple-to-measure characteristics (e.g., DBH and THT) can be evaluated for their ability to predict volume. Research has shown that one of the most useful characteristics is called the "combined variable." This is simply $DBH^2 \times THT$ for an individual tree (note: the units are not converted). Foresters then use a technique called linear regression (usually performed with software) to find the statistically "best" straight line through the points. The form of this line is $y = mx + b$, where y is the volume, m is the slope of the line, x is the combined variable ($DBH^2 \times THT$), and b is the point at which the line intersects the y-axis. Figure 3 shows the best fitting straight line using the combined variable to predict tree volume with the following formula: $pred.vol = b + m \times (DBH^2 \times THT)$.

In this example, the optimal slope (m) was found to be 0.0022, and the optimal y-intercept (b) was found to be 1.4455.

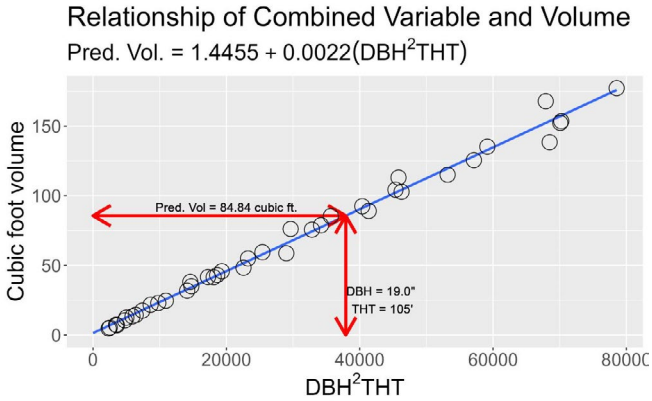

Figure 3. The relationship between the combined variable $DBH^2 \times THT$ and total cubic foot volume where DBH = diameter at breast height and THT = total height.

Use and Implications

So how is this line useful? An example is shown in figure 3. Assume you measure the DBH of a tree to be 19 inches and the total height to be 105 feet. The combined variable is $19^2 \times 105 = 37{,}905$. Using the regression equation, the predicted volume for the tree would be 84.84 $ft^3 = 1.4455 + 0.0022 \times 37{,}905$. Using two relatively easy-to-measure characteristics (DBH and THT), an extremely difficult value to measure (total volume) was predicted.

A natural question is how good is this prediction? This is an important question to ask because a statistically best line can always be produced, even with weak relationships. Compare the two graphs in figure 4.

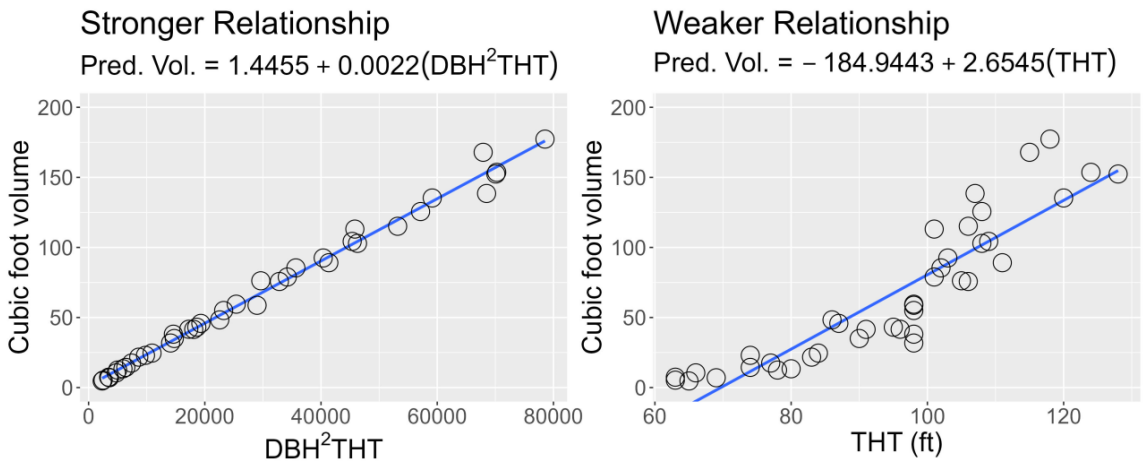

Figure 4. Comparison of two prediction equations for total tree volume where DBH = diameter at breast height and THT = total height.

By looking at the spread of points around the line, it's clear that the line in the left panel of figure 4 is a better fit to the data.

Software can calculate two common values that can be used to determine how good the prediction equation is. The first is called "R-squared." This essentially tells us what percentage of the variation in the y-characteristic the regression line explains. The value ranges from 0 to 1, with 1 being perfect and 0 indicating no relationship.

Another value used is the standard error of the estimate. This is a measure of a typical deviation from the regression line. In other words, what is a typical deviation above or below the regression line of measured volume compared to the value predicted by the line? This can be any value greater than or equal to zero and is expressed in the same units as the value being predicted. Smaller values are preferred (i.e., a smaller typical deviation is better). Table 1 shows the statistics for the two different regression equations.

Table 1. Fit statistics for the two predictor variables considered where DBH = diameter at breast height and THT = total height.

Predictor (x) used	R-squared	Standard error (ft^3)
DBH^2 x THT	99%	4.84
THT	82%	21.92

The statistics confirm our visual interpretation that using the combined variable is preferred. The R-squared value is closer to 1, and the standard error is much lower.

When using regression equations, there are a few important risks to be aware of. First, be careful when using the equation to predict far outside the range of data it was produced with (e.g., DBH, THT, species, location). This is called "extrapolation" and can lead to nonsensical results. Consider the equation if using only height to predict volume. For smaller trees, the predicted volume will actually end up negative! This equation can safely be used for the range of data it was produced from, but extrapolation led to major issues.

Another important point to consider is if a straight line is appropriate. Consider figure 5.

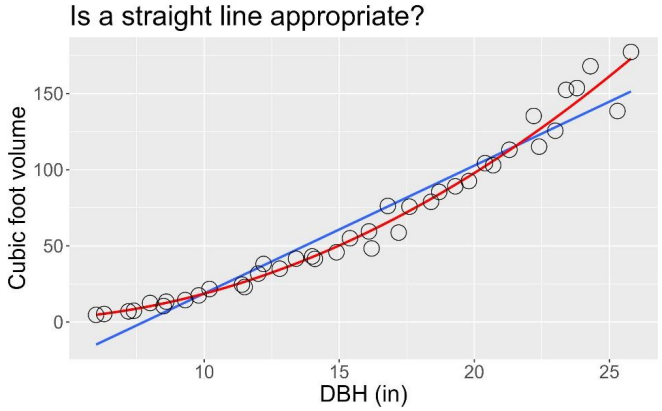

Figure 5. Using DBH alone to predict total tree volume results in a relationship that is not best described by a straight line (the blue line) where DBH = diameter at breast height.

In figure 5, it's clear that the relationship between DBH and cubic foot volume is best described by using a curved line rather than a straight line. The red curve is produced using a more advanced statistical technique, but it illustrates an important point: When looking at the data, study it carefully to determine if a straight line is appropriate. If not, either proceed with caution or consider a different statistical technique.

Finally, these models assume the predictor variables ($DBH^2 \times THT$) are measured without any error. The approach described has no ability to help correct for this. High-quality data are important, and procedures should be implemented to check data prior to use.

Once produced, regression equations are simple to use and can easily be programmed into a calculator or spreadsheet. Volume tables can be easily produced. For example, the following table (table 2) was produced for yellow poplar using the combined variable equation in figure 4. The advantage of the equation is the volume can be predicted for any size tree. A table for all possible DBH-THT combinations would be impractical and hard to use.

There are many extensions to the linear regression technique outlined. Multiple x-values can be used to produce equations with a method called "multiple regression." It is difficult to visualize, but using the equation is very similar. For example, DBH and THT could be used to produce the following equation:
$pred.vol = -75.8235 + (7.8355 \times DBH) + (0.2047 \times THT)$.

Notice they are not in the combined variable form. To use this prediction equation, plug in the DBH and THT measurements and solve:
$94.54\ ft^3 = -75.8235 + (7.8355 \times 19) + (0.2047 \times 105)$.

Note: This is quite different than the combined variable answer. They are not expected to be the same, and in this case, the combined variable form is statistically better (R-squared = 99% vs. 94%, and standard error = 4.84 ft^3 vs. 12.53 ft^3).

The red curve in figure 5 was produced using a more advanced statistical procedure called non-linear regression; however, using the equation is similar:
$pred.vol = 4.0056 - 2.1161 \times DBH^{1.3505} + 0.9941 \times DBH^{1.7944}$.

$87.05\ ft^3 = 4.0056 - 2.1161 \times 19^{1.3505} + 0.9941 \times 19^{1.7944}$.

In summary, the regression technique is a highly useful method to predict characteristics that are difficult to measure. Regression has many applications throughout forestry, and software can easily produce and implement equations. Many scientific studies have developed highly accurate and useful predictive equations for many tree characteristics. This leads to more realistic estimates of, for example, timber value, carbon storage, and wildlife habitat. However, before using any regression equation, be sure to examine the range of data used to produce it to guard against illogical predictions due to extrapolation. Also, review any fit statistics (e.g., R-squared and standard error) to determine if the regression equation fits the data well, so you have confidence in the predictions.

Table 2. Volume table for yellow poplar produced with the prediction equation: $pred.vol = 1.4455 + 0.0022 \times (DBH^2 \times THT)$ where DBH = diameter at breast height and THT = total height.

	THT (ft)								
DBH (in)	50	60	70	80	90	100	110	120	130
6	5.41	6.20							
8	8.49	9.89	11.30						
10	12.45	14.65	16.85	19.05					
12		20.45	23.62	26.79	29.96	33.13			
14			31.63	35.94	40.25	44.57			
16				46.50	52.13	57.77			
18					65.60	72.73	79.85		
20						89.45	98.25	107.05	
22						107.93	118.57	129.22	139.87
24						128.17	140.84	153.51	166.18
26						150.17	165.04	179.91	194.78
28						173.93	191.17	208.42	225.67
30						199.45	219.25	239.05	258.85

References

Burkhart, Harold E., Thomas E. Avery, and Bronson P. Bullock. 2019. *Forest Measurements*, 6th ed. Long Grove, IL: Waveland Press.

Visit our website: www.ext.vt.edu
Produced by Virginia Cooperative Extension, Virginia Tech

VT/1024/CNRE-184P