Discussion Solution — Height of *Halimeda* in Relation to Depth

Is mean height related to depth? If so, how?

Summary

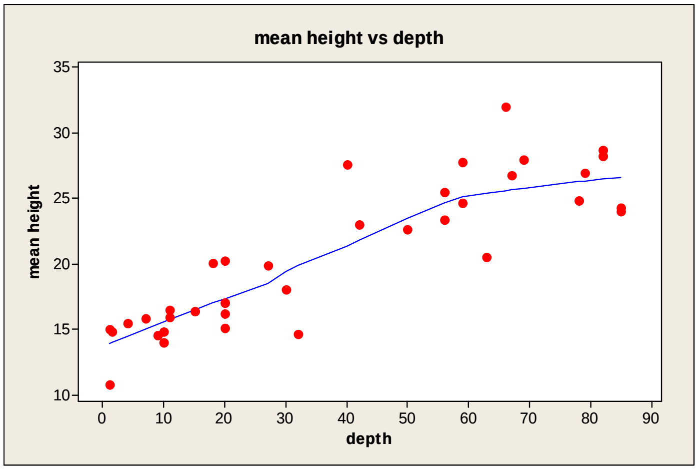
Linear least-squares regression analysis yields the conclusion that **the mean height of *Halimeda* increases with increasing depth, by approximately 17 cm / m** (95% CI: 0.137 – 0.201; P-value < 0.0005). The relationship is mildly non-linear (less steep at greater depths) and the variability is mildly uneven (less at shallow depths), but these violations of regression assumptions are not severe enough to invalidate the analysis.

Alternatively, correlation analysis could be used, but parametric inference about ρ using r is not valid because the distributions (particularly of depths) are very non-Normal. Spearman's rank correlation and a randomization test both produce the similar conclusion that height and depth are positively associated, with P-values < 0.0005.

Initial Considerations

This study deals with the relationship between two quantitative variables, so the analysis should be by regression or correlation or something similar. It would be natural to think of depth as the explanatory variable and height the response, suggesting a regression approach, but a fitted regression model is not essential to answering the question so correlation analysis might be adequate.

A scatterplot shows a fairly steep and tight positive association, with typical heights more than doubling from the shallowest to the deepest sites. The pattern is mostly straight but possibly less steep at the upper-right. There also is more variability in height at the greater depths.



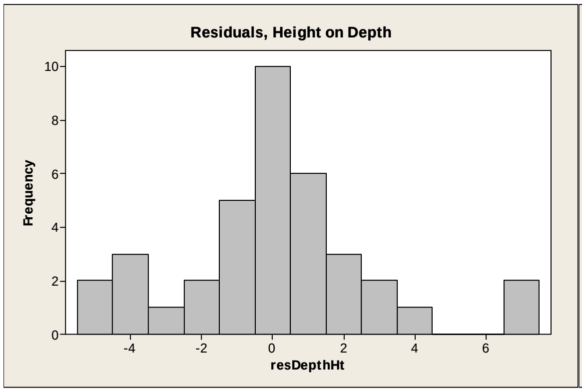
Regression

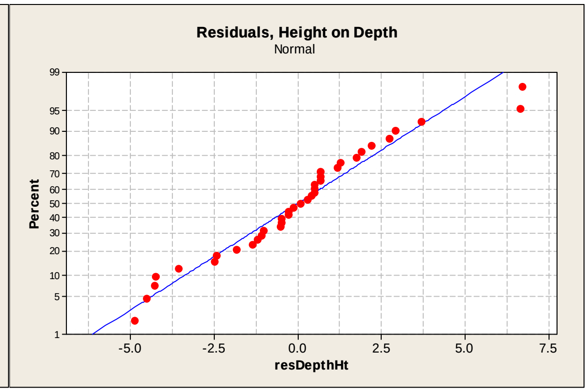
Assumptions

Although the sites were not randomly sampled, it seems reasonable to consider the data as representative of *Halimeda* at the various depths. It is possible, perhaps likely, that nearby sites, or sites in similar environments, might be more similar to one another, indicated a kind of non-independence. If we knew the locations of the sites it might be useful to include location information in a more complex model. Lacking this information, though, I think it is reasonable to treat these observations as independent; certainly knowing the value of any observation doesn't give us any information about any other observation other than in relation to depth, our explanatory variable.

The depths of the sites presumably are measured fairly precisely, at least relative to the variation in depth among the sites, so least squares regression (*i.e.* assuming values of the explanatory variable are known) seems reasonable.

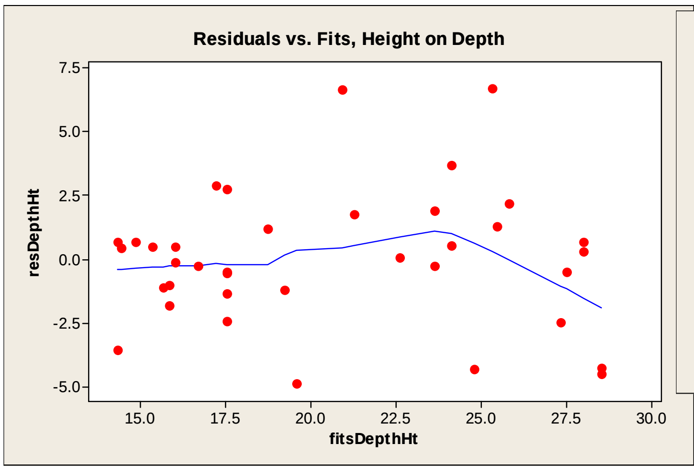
Linear least squares regression therefore is at least provisionally acceptable for these data; further assessment can be done using regression residuals. The distribution of the residuals (below) has several moderately non-Normal aspects: the middle of the distribution is "peaky" (a high peak and then stretched tails), but then there is a lump — a small second peak — at the bottom end. Probably most important are two high outliers.

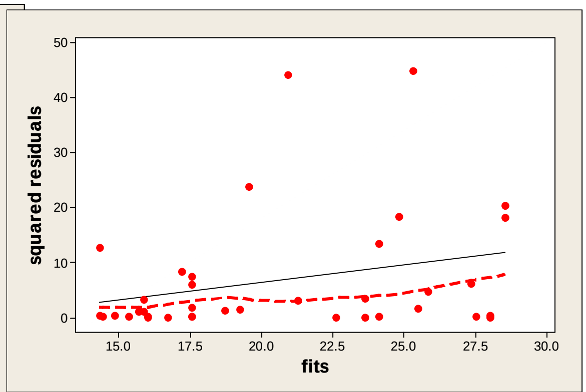




The plot of residuals *vs*. fits (below, left panel) shows a little more clearly two issues noted in the original scatter plot: the curvature in the trend at the right, and the greater variability at the right.

An additional way to visualize the evenness or unevenness of the variability is to plot the squared residuals (or absolute values of residuals) against the fits or the explanatory variable. Such a plot is in the right panel below, and shows a clear trend of increasing variability (larger residuals) as the fit (or depth) increases. This trend of increasing spread, though, is not steep, nor is the nonlinearity of the relationship of height to depth severe, so I feel the least squares regression is acceptable as at least an approximate description of the relationship.





Transformation?

Our primary tool for dealing with nonlinear trends and/or uneven variability is transformation. For these data, however, transforming the mean heights to reduce the variability at high values relative to that at lower values will also increase the curvature at the upper end of the trend, so a transformation of the depths would also be needed to compensate for this.

I examined various combinations of transformations of the two variables, but all things considered, I do not feel any transformations improved the agreement of the data with the assumptions of regression sufficiently to prefer them over analysis of the original data. I therefore would answer this question using linear least-squares regression on the untransformed data.

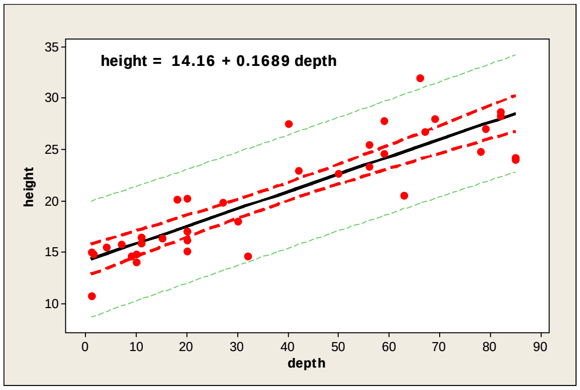
Lacking any reason to expect the alga to be either shorter or taller at greater depths, a two-sided test is appropriate, with the null hypothesis being that mean height is not related to depth (the slope is 0).

The test is highly significant and the model explains more than 3/4ths of the variance in mean height:

Predictor	Coef	SE Coef	T	P
Constant	14.1565	0.7398	19.14	0.000
depth	0.16889	0.01565	10.79	**0.000**

S = 2.67859   **R-Sq = 76.9%**

The 95% CI for the slope is 0.16889 ± 2.03 × 0.01565 = (0.1371, 0.2007). As both this narrow CI and the graph at right of the 95% confidence bounds for the mean height as a function of depth show, the data provide a fairly precise estimate of the mean relationship between the two variables, though there is a moderate amount of variation among site mean depths for any given depth.

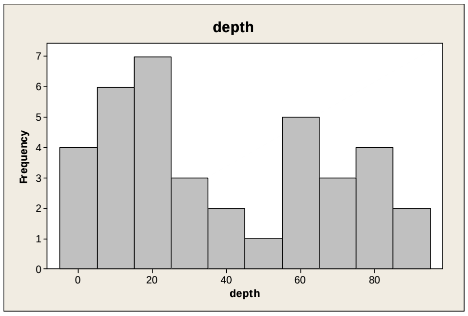


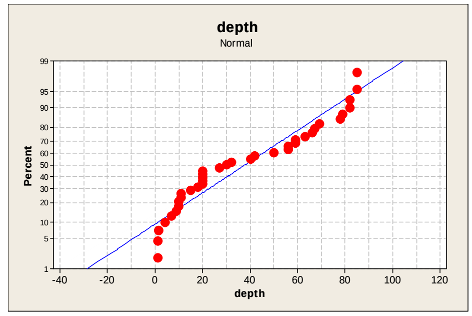
Correlation

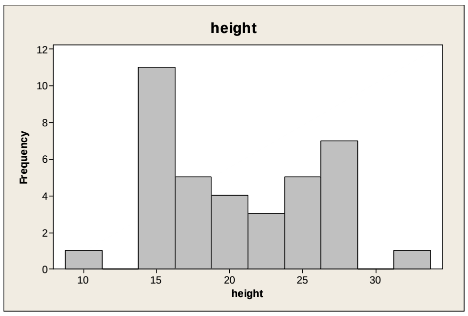
Alternatively, correlation analysis might be prefered, on the grounds that the depth variable does have some measurement error and the regression equation is not needed. The data do not, however, meet the assumption of bivariate Normality: although the shape of the scatterplot is acceptable, neither variable is Normally distributed, as shown at the top of the next page. This is the result primarily of the sites having been chosen to give a range of depths, and with a preference for similar and round-numbered depths to facilitate comparison among sites (which was not part of our current analysis).

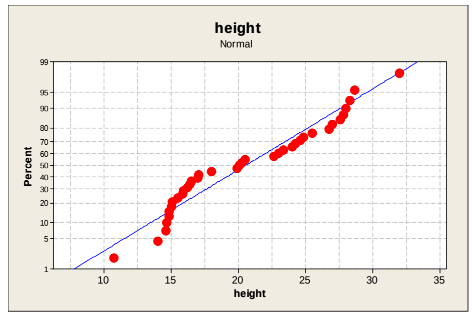
Because of the clear non-Normality, combined with the mild non-linearity and uneven variance, a randomization test of correlation, or Spearman's rank correlation, would be reasonable choices. Both test the null hypothesis of no association. The randomization test uses Pearson's correlation coefficient as the test statistic and so tests for a linear association, but without assuming bivariate Normality; Spearman's rank correlation tests for linear correlation of the ranks and thus for a monotonic association between the variables.

Both correlation tests are highly significant, as shown on the next page below the graphs.









<u>Spearman's:</u>

```
Pearson correlation of rankDepth and rankHeight = 0.866
P-Value = 0.000
```

<u>randomization:</u>

```
                          Observed      Mean      SE alternative p-value
cor(depth,height)     0.8769 0.0001451 0.1684   two.sided  0.0002
```

(The randomization test in fact would give a much smaller *P*-value with more randomizations. The result shown above was from 9999 randomizations, in which the largest randomized correlation coefficient was less than 0.6, <u>far</u> smaller than the observed r = 0.8769.)

Choice of Method, and Conclusion

I would use the regression analysis, and conclude that **yes, mean height of *Halimeda* increases significantly and approximately linearly with increasing depth, by about 0.17 cm per m**. A similar but unquantified conclusion would come from either of the non-Pearson correlation analyses: **yes, mean height of *Halimeda* increases significantly and approximately linearly with increasing depth**.