Discussion # 3 — Stinkbug Age Pigment: Solution

Summary

Age is the explanatory variable and pigment the response, due to the way the study was conducted. A scatterplot indicates a roughly linear positive relationship, with considerable variability in pigment at a given age and some suggestion of curvature, that the rate of increase in pigment is less at greater ages.

Least-squares linear regression gives the estimated regression

\[ \hat{y} = 3.928 + 1.67 \cdot x \]

with \( R^2 = 0.696 \).

Residual plots show the slight nonlinearity but no other major problems. Square-root transforming the age variable slightly improves the linearity.

Using the linear equation above as an acceptable approximation to the relationship, plugging in the pigment levels of 9 and 15 for \( y \) and solving for \( x \) gives estimated ages of 3.04 and 6.63 weeks.

Details

This question is about the relationship between two quantitative variables. Since the data are from an experiment in which the age variable was controlled (and known exactly), and the resulting pigment amounts were observed, the proper analysis treats age as the explanatory variable, despite the fact that the intended future use is to estimate bug ages from pigment levels. To be able to do this “inverse prediction” an explicit numerical relationship between the two variables is needed, not simply a measure of their association. Regression therefore is the appropriate statistical tool, so long as it is a reasonable description of the observed relationship.

A scatterplot, possibly with smoother (see next page), shows a generally linear increase in pigment level with increasing age, and no obvious reason not to perform a regression. There is slight curvature (pigment accumulation slowing with age?), there are a few somewhat unusual observations (e.g., the highest pigment level at age = 2), and the spread of the points varies somewhat among the ages (standard deviations range from 1.11 to 2.43), but these problems are not severe.

Running the regression and examining residuals, though, makes the curvature seem more of a problem: the residuals tend to be negative at high and low values of age and positive at intermediate values (this can be seen by putting a smoother on a plot of residuals vs. age or vs. fits, as shown on the next page). This suggests that the relationship is not quite linear, but curving off slightly to the right (decelerating, concave downward). The residuals are slightly skewed (seen better in a histogram [next page] than a NQQ plot), but not enough to cause any concern.

The curvature could be dealt with in several ways, some of which (nonlinear or polynomial regression) are beyond the scope of this course. One possible solution which is within the scope of this course would be to transform the time variable to pull the top-right part of the relationship slightly to the left; either square-root or log transformation works reasonably well for linearizing the relationship.
Another way to deal with the curvature would be simply to decide that it is slight enough that it doesn’t need to be dealt with, that a linear regression is a useful approximation to the relationship. Taking this approach, the estimated regression line is

\[ \hat{y} = 3.928 + 1.67 \cdot x \]

with \( R^2 = 0.696 \).

Clearly an estimation of age from pigment level will not be very precise, since a wide range of ages could have a given pigment level (and vice versa). However, it is reasonable to use the regression model to get a point estimate of the age for a particular pigment level. The regression model rearranges to

\[ \hat{x} = (y^* - 3.928) / 1.67 \]

giving ages of 3.04 and 6.63 weeks for pigments levels of 9 and 15. Note that the latter estimate involves extrapolating slightly beyond the data, which always is risky.
Any reasonable non-straight model for the relationship — polynomial regression, regression on transformed age, or simply the smoother of the original data — would yield a lower estimate of the age for a pigment level of 9, and a higher estimate for a level of 15 (the latter now requiring substantial extrapolation beyond the data).

Since the intent is to use pigment levels to estimate ages it is tempting to treat pigment as the response variable. As noted above, however, this is not appropriate since there is both measurement error and ‘random’ (i.e. real but unexplained) variation among bugs in the pigment level while the ages were controlled and known quite precisely. The least-squares criterion for fitting the line, which tries to minimize variability around the line in the vertical direction therefore would not be appropriate for a regression of age on pigment. The same $R^2$ of course is obtained regardless of which way the regression is run, but how the line fits the data will differ, and the estimates of age for a given pigment level therefore also will differ.