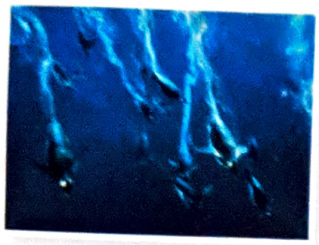
WHAT HAVE WE LEARNED?



We've learned to be alert to the many ways in which a data set may be unsuitable for a regression analysis.

- Watch out for more than one group hiding in your regression analysis. If you find subsets of the data that behave differently, consider fitting a different regression model to each subset.
- The Straight Enough Condition says that the relationship should be reasonably straight to fit a regression line. Somewhat paradoxically, sometimes it's easier to see that the relationship is not straight *after* fitting the regression line by examining the residuals. The same is true of outliers.
- The Outlier Condition actually means two things: Points with large residuals or high leverage (especially both) can influence the regression model significantly. It's a good idea to perform the regression analysis with and without such points to see their impact.

And we've learned that even a good regression model doesn't mean we should believe that the model says more than it really does.

- Extrapolation far from $\bar{x}$ can lead to silly and useless predictions.
- Even an R^2 near 100% doesn't indicate that x causes y (or the other way around). Watch out for lurking variables that may affect both x and y.
- Be careful when you interpret regressions based on *summaries* of the data sets. These regressions tend to look stronger than the regression based on all the individual data.

TERMS

Extrapolation Although linear models provide an easy way to predict values of y for a given value of x, it is unsafe to predict for values of x far from the ones used to find the linear model equation. Such extrapolation may pretend to see into the future, but the predictions should not be trusted. (p. 207)

Outlier Any data point that stands away from the others can be called an outlier. In regression, outliers can be extraordinary in two ways: by having a large residual or by having high leverage. (p. 210)

Leverage Data points whose x-values are far from the mean of x are said to exert leverage on a linear model. High-leverage points pull the line close to them, and so they can have a large effect on the line, sometimes very strongly influencing the slope and intercept. With high enough leverage, their residuals can be deceptively small. (p. 211)

Influential point If omitting a point from the data results in a regression model with a very different slope, then that point is called an influential point. (p. 211)

Lurking variable A variable that is not explicitly part of a model but affects the way the variables in the model appear to be related is called a lurking variable. Because we can never be certain that observational data are not hiding a lurking variable that influences both x and y, it is never safe to conclude that a linear model demonstrates a causal relationship, no matter how strong the linear association. (p. 212)

ON THE COMPUTER

Regression Diagnosis

Most statistics technology offers simple ways to check whether your data satisfy the con[illegible]egress[illegible]
We have already seen that these programs can make a simple scatterplot. They can als[illegible]ck th[illegible]
conditions by plotting residuals.