Chapter 9 - Regression Wisdom

August 24, 2010

Regression is one of more widely used statistical methods. It is also widely abused and misinterpreted.

Make sure association is linear before finding the Correlation and the Linear Regression Line. Sometimes it is hard to tell that a relationship is nonlinear by the Scatterplot, but the Residual plot will show the "curve" better.

Duration of Dive and Dive Heart Rate for Emperor Penquins. $(R^2 = 71.5\%)$

Fig 9.1, p. 228

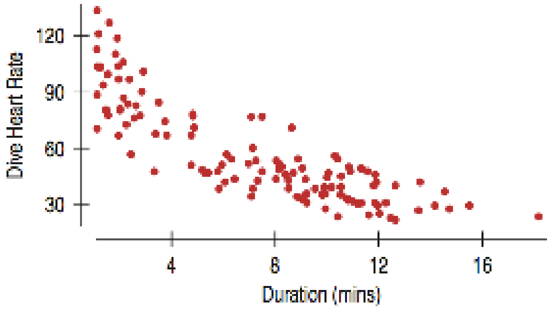
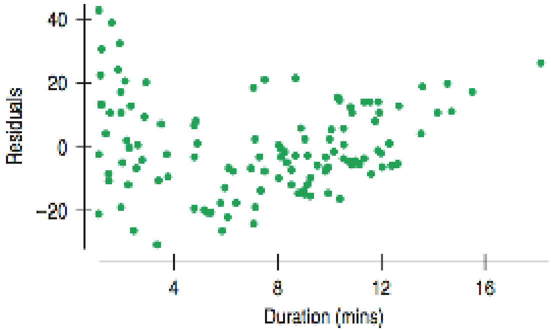


Fig 9.2, p. 228



Sifting Residuals for Groups

In Chapter 8 the book looked at an example where the relationship between sugar content and calories of cereals were compared.

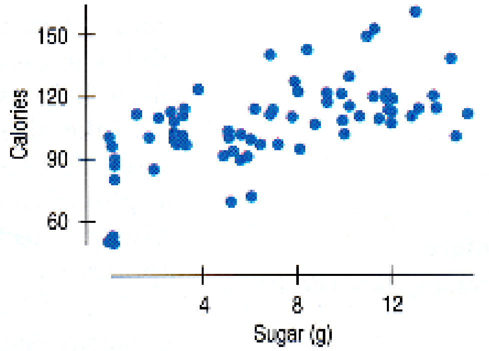


Fig. 9.3, p. 229

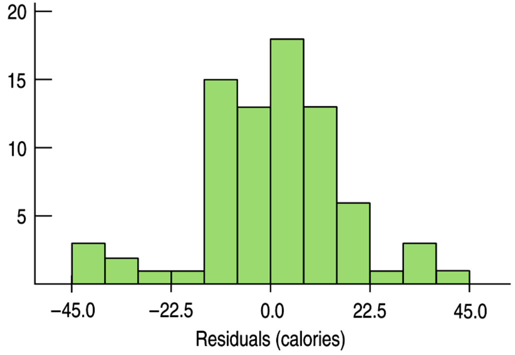
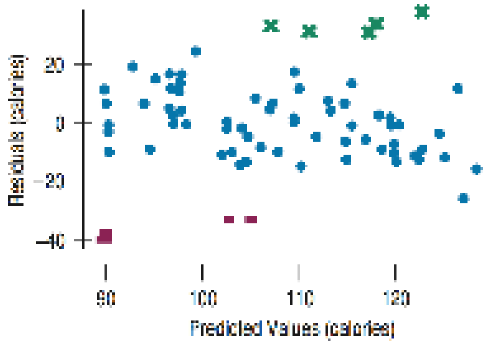
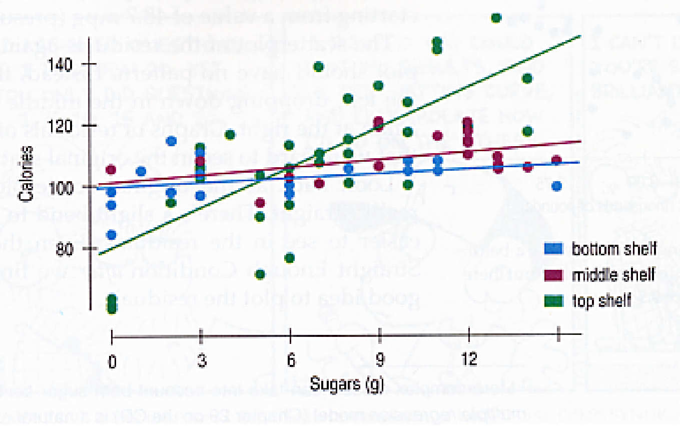


Fig. 9.4, p. 229



Watch out for Subsets in the data. The data must come from the same group.

Fig 9.5, p. 230



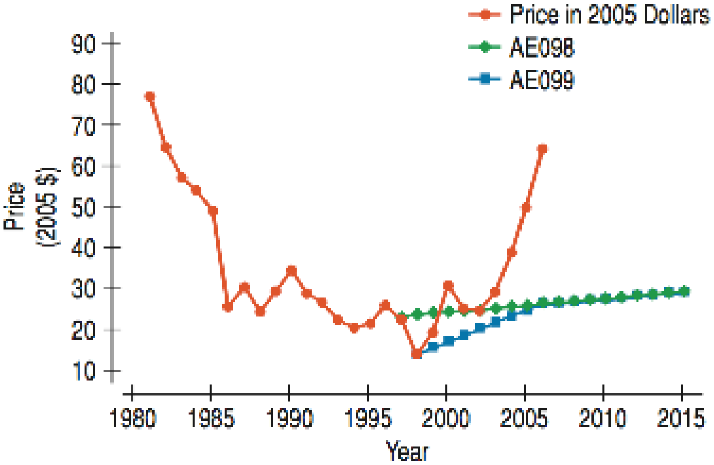
Cereals tend to placed at the eye level of the proposed consumer. Cereals for kids tend to be on the lower shelves while cereals for adults tend to be on the higher shelves.

Do you report the regression separately for cereals on the top shelf and cereals on the bottom two shelves?

Watch out for Extrapolation. Extrapolation is when the model is used to predict for values of *x* not in the range of the data. If *x* variable is time then extrapolation is an attempt to see into the future.

Book discusses oil prices. Model from 1971 to 1982 has prices increasing at a rate of $7 per year. Model from 1981 to 1998 shows prices decreased at a rate of $3 per year. Then after 1998 to 2006 there was a sharp increase.

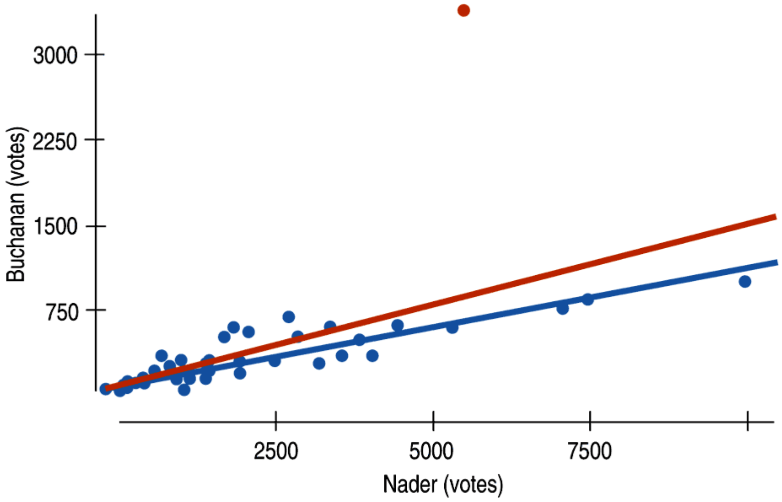
A timeplot from 1981 to 2006 of actual prices along with the Energy Information Administration (EIA) predictions is below. Fig. 9.8, p. 232



If extrapolation must be used do not believe it will be correct.

Watch out for Outliers.
The following scatterplot shows the number of Nader votes vs number of Buchanan votes for each Florida County in the 2000 Presidential Election (Bush vs Gore).

Fig 9.10, p. 233



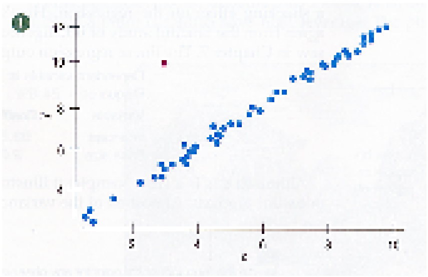
Blue line has slope 0.14 and $R^2 = 42.8\%$. Red line has slope 0.1 and $R^2 = 82.1\%$

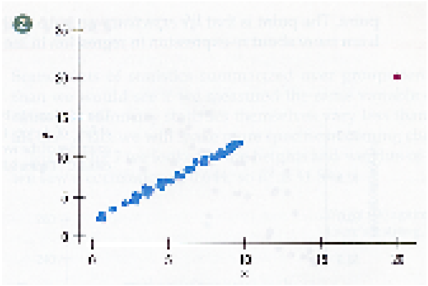
An **Outlier** is any point that stands away from the others.

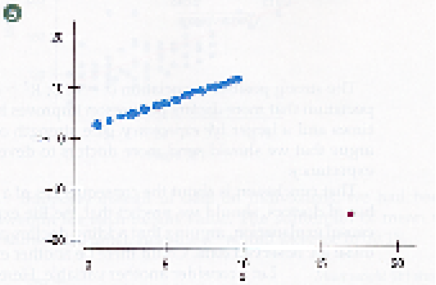
- Points with **Large Residuals** are far from the regression line.

- Points with **High Leverage** pull the regression line towards them altering the slope and intercept. These points have x-values that are far from $\bar{x}$, think of $\bar{x}$ as being the balancing point for a lever. Their residuals may appear to be small.

A point is considered to be **influential** if omitting the point gives a very different model.

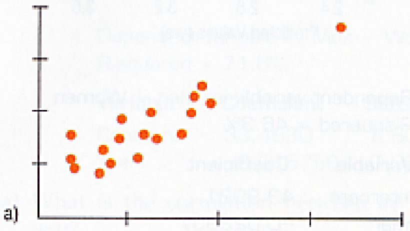


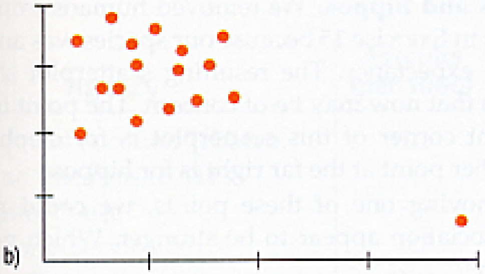


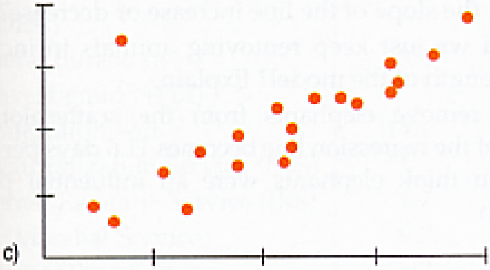


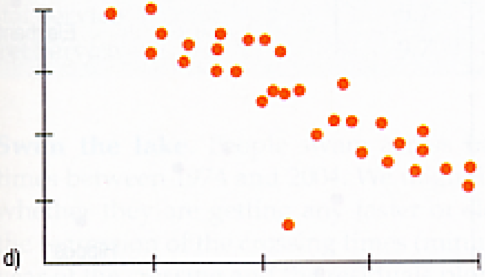
11. **Unusual points.** Each of the scatterplots below shows a cluster of points and one "stray" point. For each, answer these questions:

1) In what way is the point unusual? Does it have high leverage, a large residual, or both?
2) Do you think that point is an influential point?
3) If that point were removed from the data, would the correlation become stronger or weaker? Explain.
4) If that point were removed from the data, would the slope of the regression line increase or decrease? Explain.



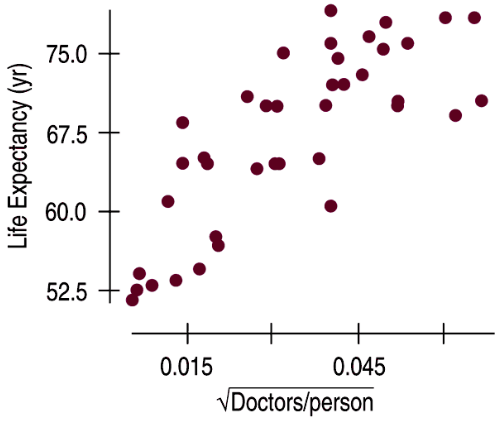






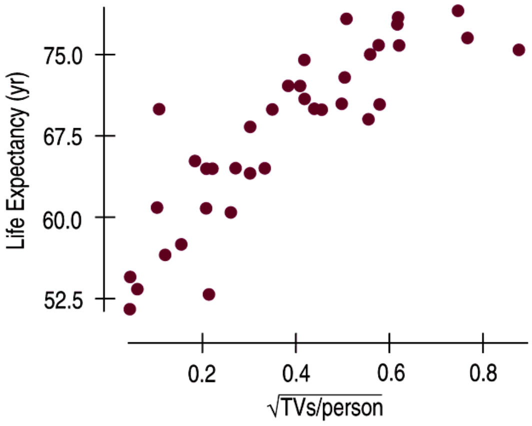
Beware of Lurking Variables. Lurking Variable is a variable that affects the way the variables in the model appear to be related. No matter how strong the correlation between two variable is this does not give causation (x causes y), because lurking variables can never be ruled out.

Fig 9.13, p. 236



$R^2 = 62.4\%$

Fig 9.14, p. 237



$R^2 = 72.3\%$

Watch out for data that involves summary values. Summary values like mean and median will not have the variability as actual data values from individuals.