Chapter 9: Regression Wisdom

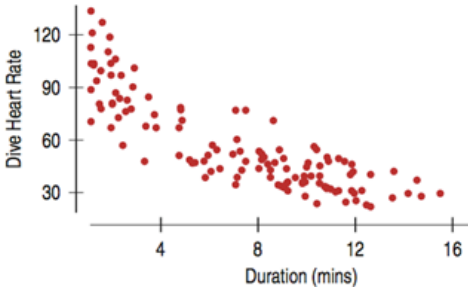
Obj - SWBAT include diagnostic information such as plots of residuals and leverages as part of your report of a regression.

Getting the "Bends"

- Linear regression only works for linear models. (That sounds obvious, but when you fit a regression, you can't take it for granted.)
- A curved relationship between two variables might not be apparent when looking at a scatterplot alone, but will be more obvious in a plot of the residuals.
- Remember, we want to see "nothing" in a plot of the residuals.

Getting the “Bends” (cont.)

- The scatterplot of residuals against Duration of emperor penguin dives holds a surprise. The Linearity Assumption says we should not see a pattern, but instead there is a bend.
- Even though it means checking the Straight Enough Condition after you find the regression, it’s always good to check your scatterplot of the residuals for bends that you might have overlooked in the original scatterplot.

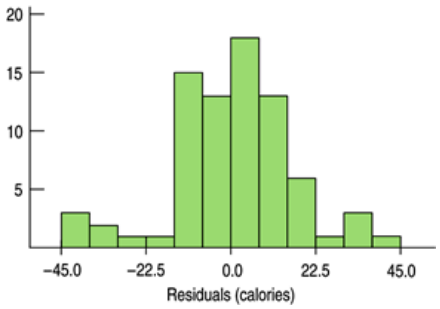


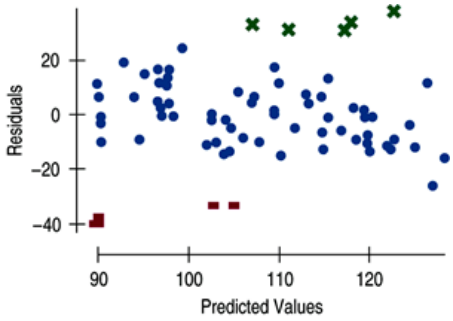
Sifting Residuals for Groups

- No regression analysis is complete without a display of the residuals to check that the linear model is reasonable.
- Residuals often reveal subtleties that were not clear from a plot of the original data.

Sifting Residuals for Groups (cont.)

- It is a good idea to look at both a histogram of the residuals and a scatterplot of the residuals vs. predicted values in the regression predicting Calories from Sugar content in cereals:





- The small modes in the histogram are marked with different colors and symbols in the residual plot above. What do you see?

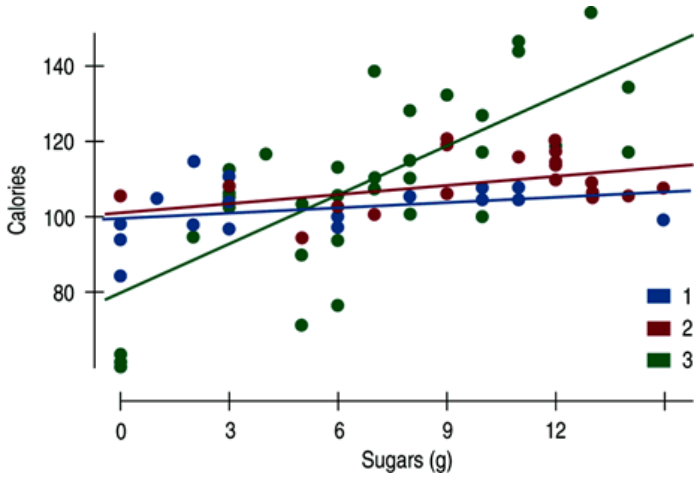
Subsets

- Here's an important unstated condition for fitting models:

- When we discover that there is more than one group in a regression, neither modeling the groups together nor modeling them apart is necessarily correct. You must determine what makes the most sense.

Subsets (cont.)

- The figure shows regression lines fit to calories and sugar for each of the three cereal shelves in a supermarket:



Extrapolation: Reaching Beyond the Data

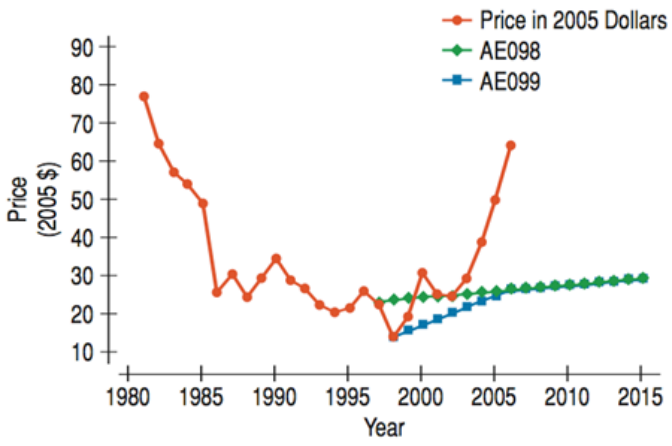
- Linear models give a predicted value for each case in the data.
- We cannot assume that a linear relationship in the data exists beyond the range of the data.
- The farther the new x value is from the mean in x, the less trust we should place in the predicted value.

Extrapolation (cont.)

- Extrapolations are dubious because they require the additional—and very questionable — assumption that nothing about the relationship between x and y changes even at extreme values of x.
- Extrapolations can get you into deep trouble. You're better off not making extrapolations.

Extrapolation (cont.)

- Here is a timeplot of the Energy Information Administration (EIA) predictions and actual prices of oil barrel prices. How did forecasters do?



- They seemed to have missed a sharp run-up in oil prices in the past few years.

Predicting the Future

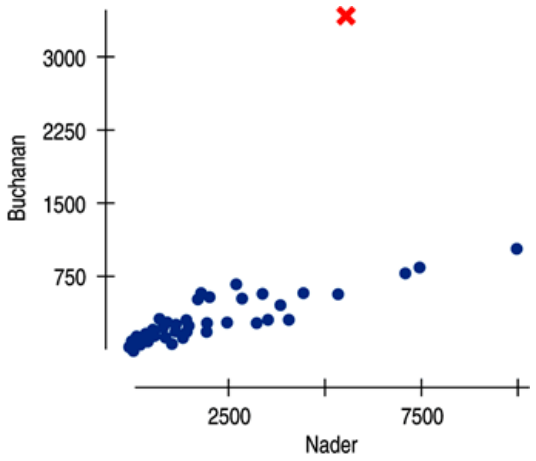
- Extrapolation is always dangerous. But, when the x-variable in the model is time, extrapolation becomes an attempt to peer into the future.
- Knowing that extrapolation is dangerous doesn't stop people. The temptation to see into the future is hard to resist.
- Here's some more realistic advice: If you must extrapolate into the future, at least don't believe that the prediction will come true.

Outliers, Leverage, and Influence

- Outlying points can strongly influence a regression. Even a single point far from the body of the data can dominate the analysis.
- Any point that stands away from the others can be called an outlier and deserves your special attention.

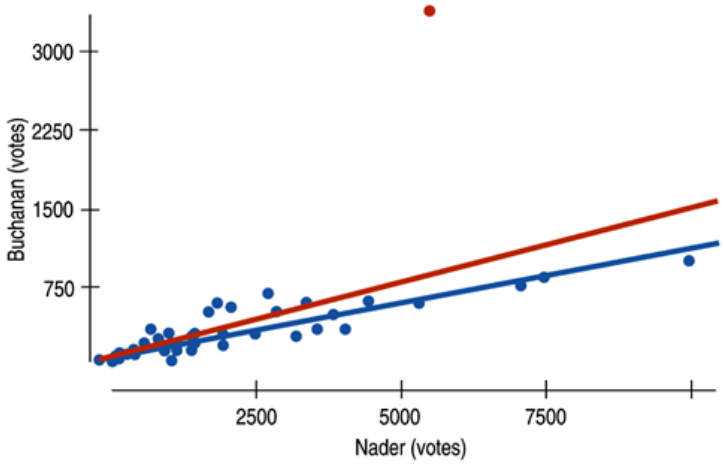
Outliers, Leverage, and Influence (cont.)

- The following scatterplot shows that something was awry in Palm Beach County, Florida, during the 2000 presidential election…



Outliers, Leverage, and Influence (cont.)

- The red line shows the effects that one unusual point can have on a regression:



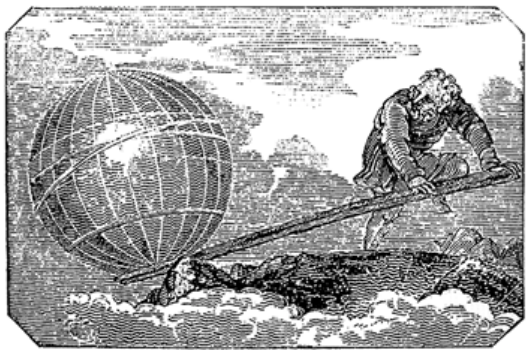
Outliers, Leverage, and Influence (cont.)

- The linear model doesn't fit points with large residuals very well.
- Because they seem to be different from the other cases, it is important to pay special attention to points with large residuals.

Outliers, Leverage, and Influence (cont.)

A data point can also be unusual if its x-value is far from the mean of the x-values. Such points are said to have high leverage.

"Give me a place to stand and I will move the Earth."
—Archimedes (287–211 BCE)

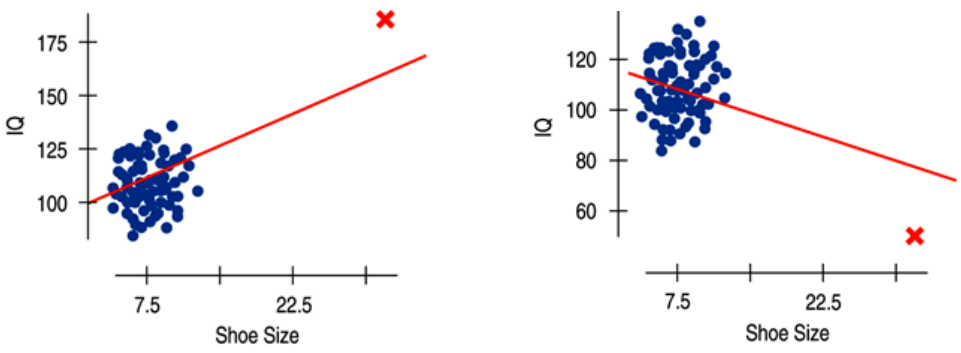


Outliers, Leverage, and Influence (cont.)

- A point with high leverage has the potential to change the regression line.
- We say that a point is influential if omitting it from the analysis gives a very different model.

Outliers, Leverage, and Influence (cont.)

The extraordinarily large shoe size gives the data point high leverage. Wherever the IQ is, the line will follow!



Outliers, Leverage, and Influence (cont.)

- When we investigate an unusual point, we often learn more about the situation than we could have learned from the model alone.
- You cannot simply delete unusual points from the data. You can, however, fit a model with and without these points as long as you examine and discuss the two regression models to understand how they differ.

Outliers, Leverage, and Influence (cont.)

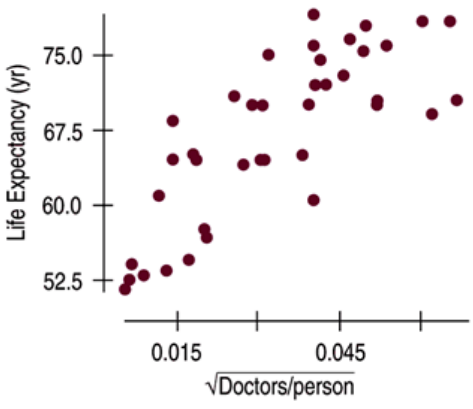
- Warning:
  - Influential points can hide in plots of residuals.
  - Points with high leverage pull the line close to them, so they often have small residuals.
  - You'll see influential points more easily in scatterplots of the original data or by finding a regression model with and without the points.

Lurking Variables and Causation

- No matter how strong the association, no matter how large the R^2 value, no matter how straight the line, there is no way to conclude from a regression alone that one variable causes the other.
- There's always the possibility that some third variable is driving both of the variables you have observed.
- With observational data, as opposed to data from a designed experiment, there is no way to be sure that a lurking variable is not the cause of any apparent association.

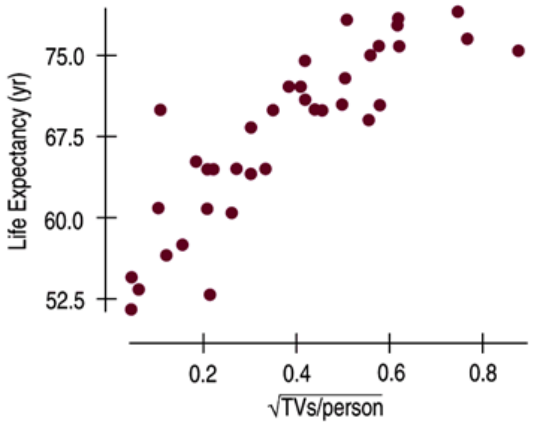
Lurking Variables and Causation (cont.)

- The following scatterplot shows that the average life expectancy for a country is related to the number of doctors per person in that country:



Lurking Variables and Causation (cont.)

- This new scatterplot shows that the average life expectancy for a country is related to the number of televisions per person in that country:



Lurking Variables and Causation (cont.)

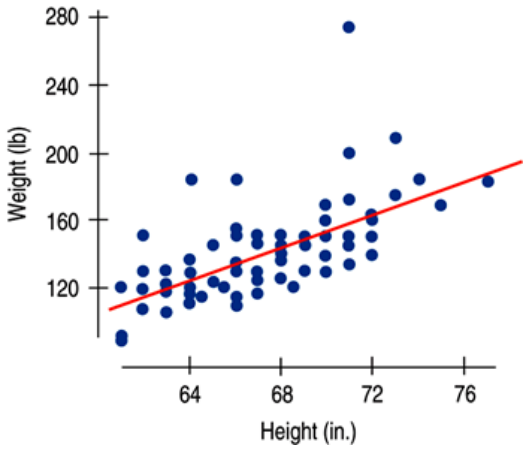
- Since televisions are cheaper than doctors, send TVs to countries with low life expectancies in order to extend lifetimes. Right?
- How about considering a lurking variable? That makes more sense…

Working With Summary Values

- Scatterplots of statistics summarized over groups tend to show less variability than we would see if we measured the same variable on individuals.
- This is because the summary statistics themselves vary less than the data on the individuals do.

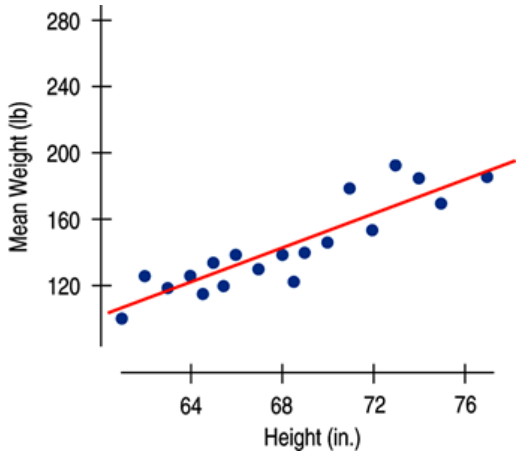
Working With Summary Values (cont.)

- There is a strong, positive, linear association between weight (in pounds) and height (in inches) for men:



Working With Summary Values (cont.)

- If instead of data on individuals we only had the mean weight for each height value, we would see an even stronger association:



Working With Summary Values (cont.)

- Means vary less than individual values.
- Scatterplots of summary statistics show less scatter than the baseline data on individuals.
- This can give a false impression of how well a line summarizes the data.
- There is no simple correction for this phenomenon.
- Once we have summary data, there's no simple way to get the original values back.

What Can Go Wrong?

- Make sure the relationship is straight.
- Check the Straight Enough Condition.
- Be on guard for different groups in your regression.
- If you find subsets that behave differently, consider fitting a different linear model to each subset.
- Beware of extrapolating.
- Beware especially of extrapolating into the future!

What Can Go Wrong? (cont.)

- Look for unusual points.
- Unusual points always deserve attention and may well reveal more about your data than the rest of the points combined.
- Beware of high leverage points, and especially those that are influential.
- Such points can alter the regression model a great deal.
- Consider comparing two regressions.
- Run regressions with extraordinary points and without and then compare the results.

What Can Go Wrong? (cont.)

- Treat unusual points honestly.
- Don't just remove unusual points to get a model that fits better.
- Beware of lurking variables—and don't assume that association is causation.
- Watch out when dealing with data that are summaries.
- Summary data tend to inflate the impression of the strength of a relationship.

What have we learned?

- There are many ways in which a data set may be unsuitable for a regression analysis:
- Watch out for subsets in the data.
- Examine the residuals to re-check the Straight Enough Condition.
- The Outlier Condition means two things:
- Points with large residuals or high leverage (especially both) can influence the regression model significantly. Perform regression analysis with and without such points to see their impact.

What have we learned? (cont.)

- Even a good regression doesn't mean we should believe the model completely:
- Extrapolation far from the mean can lead to silly and useless predictions.
- An R^2 value near 100% doesn't indicate that there is a causal relationship between x and y.
- Watch out for lurking variables.
- Watch out for regressions based on summaries of the data sets.
- These regressions tend to look stronger than the regression on the original data.

AP Statistics - Class Activity

Graduating Classes

The table at the right shows the number of seniors who graduated from a small college during the latter half of the last century. Data was not available for all of the years, especially those longer ago.

Year	Number Graduated
1950	203
1953	211
1957	288
1960	319
1962	381
1965	446
1968	439
1970	521
1972	509
1976	527
1980	476
1984	413
1987	399
1989	362
1992	379
1994	413
1996	437
1998	426
2000	451

1. Make a scatterplot and describe the trends you see in the data.

2. These data do not show the size of the graduating class in 1969. Create an appropriate model and use it to estimate the size of that class. Explain what years you used, and why.

3. If we wanted to project the graduating class sizes through the year 2010, what model would you use? What aspect of these data makes you cautious about your projections?