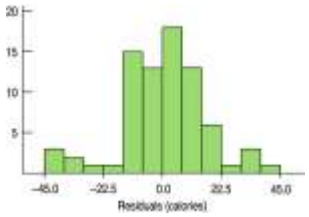
Chapter 9 Summary
Regression Wisdom

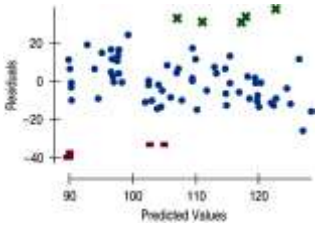
What have we learned?

- 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.
- 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.

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.
- Sometimes the subtleties we see are additional details that help confirm or refine our understanding.
- Sometimes they reveal violations of the regression conditions that require our attention.
- It is a good idea to look at both a histogram of the residuals and a scatterplot of the residuals vs. predicted values:

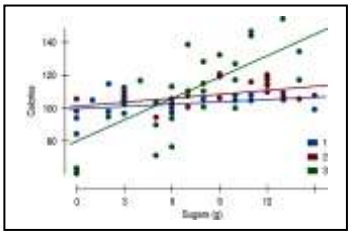




- The small modes in the histogram are marked with different colors and symbols in the residual plot above. What do you see?
- An examination of residuals often leads us to discover groups of observations that are different from the rest.
- When we discover that there is more than one group in a regression, we may decide to analyze the groups separately, using a different model for each group.

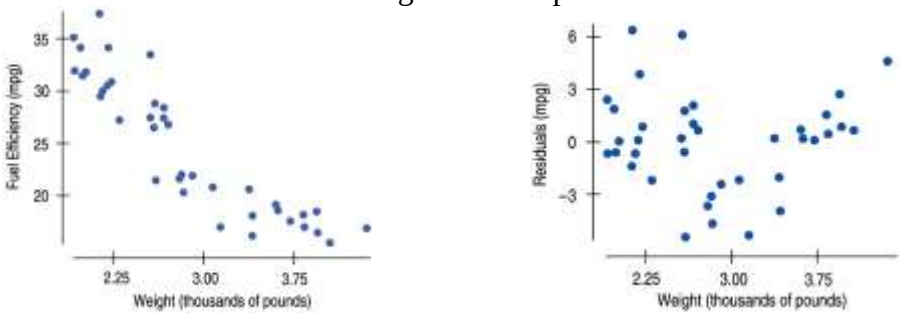
Subsets

- Here's an important unstated condition for fitting models: All the data must come from the same group.
- When we discover that there is more than one group in a regression, neither modeling the groups together nor modeling them apart is correct.
- Figure 9.3 from the text shows regression lines fit to calories and sugar for each of the three cereal shelves in a supermarket:



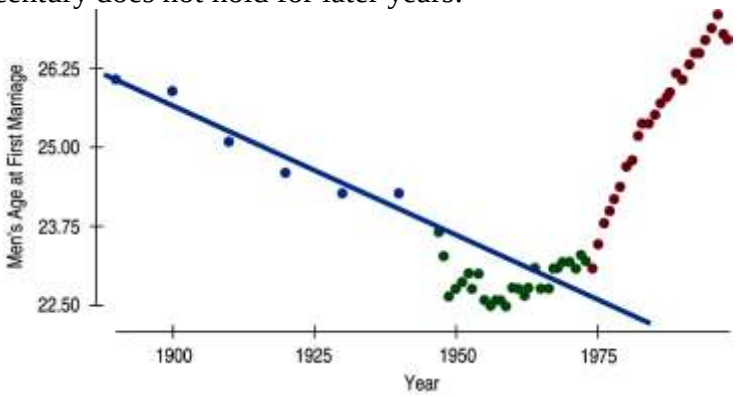
Getting the "Bends"

- 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.
- The curved relationship between fuel efficiency and weight is more obvious in the plot of the residuals than in the original scatterplot:



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.
- Once we venture into new *x* territory, such a prediction is called an extrapolation.
- 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.
- A regression of mean age at first marriage for men vs. year fit to the first 4 decades of the 20th century does not hold for later years:

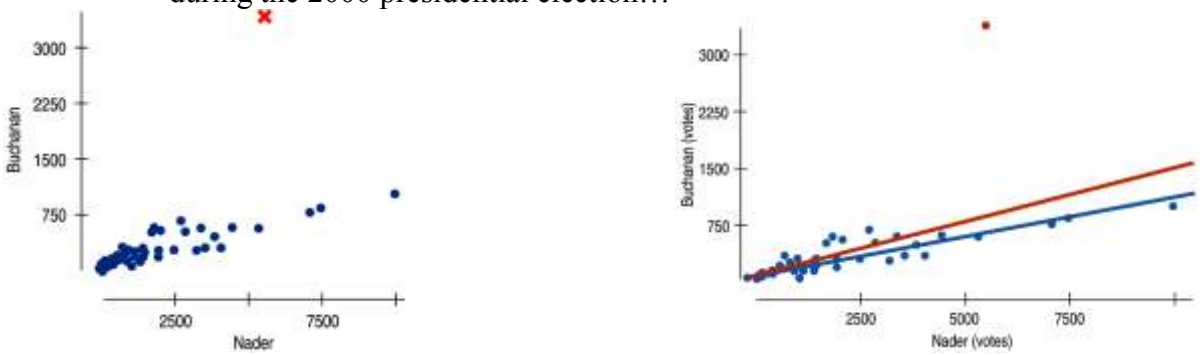


Predicting the Future

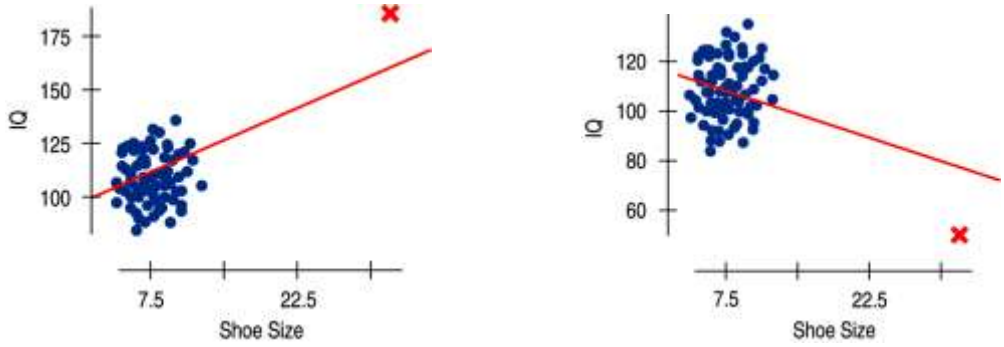
- Extrapolation is always dangerous. But, when the *x*-variable in the model is *time*, extrapolation becomes an attempt to peer into the future.
- Knowing that extrapolation is dangerous doesn't stop people. The temptation to see into the future is hard to resist.
- Here's some more realistic advice: 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.
- The following scatterplot shows that something was awry in Palm Beach County, Florida, during the 2000 presidential election…



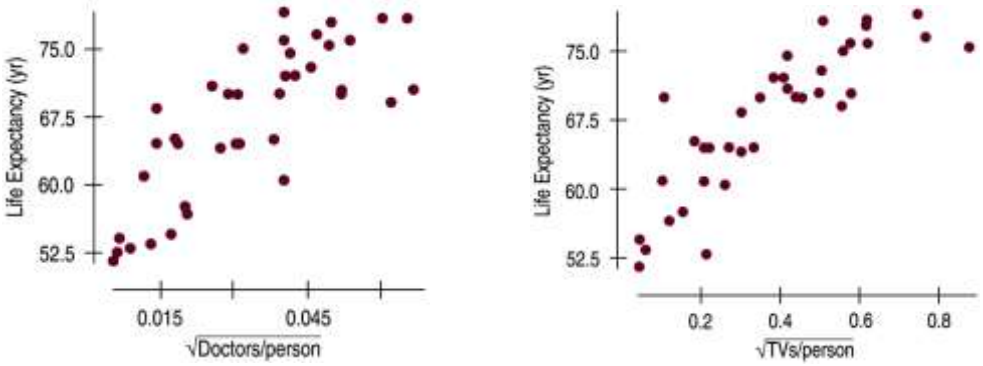
- The red line shows the effects that one unusual point can have on a regression:
- The linear model doesn't fit points with large residuals very well.
- Because they seem to be different from the other cases, it is important to pay special attention to points with large residuals.
- A data point can also be unusual if its *x*-value is far from the mean of the *x*-values. Such points are said to have high leverage.
- 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.



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



- This new scatterplot shows that the average *life expectancy* for a country is related to the number of *televisions* per person in that country:
- 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…
  - Countries with higher standards of living have both longer life expectancies *and* more doctors (and TVs!).
  - If higher living standards *cause* changes in these other variables, improving living standards might be expected to prolong lives and increase the numbers of doctors and TVs.

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.
- 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!
- Look for unusual points.
  - Unusual points always deserve attention and that 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.
- 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.