

Chapter 9: Regression Wisdom

Learning Objectives

At the end of this chapter, students will be able to:

- understand that sometimes there may be subsets in the data worth exploring separately.
- describe how unusual data points affect the regression model and the correlation coefficient.
- create a residual plot and look for patterns in the plot.

We will look at:

- pattern changes in scatterplots; the dangers of extrapolation; the possible effects of outliers, high leverage, and influential points; the problem of regression of summary data; and the mistake of inferring causation.

Chapter 9: Regression Wisdom

Issues and Problems with Regression

- Subsets and curves
- Dangers of extrapolation
- Possible effects of outliers, high leverage, and influential points
- Problems with regression of summary data
- Mistakes of inferring causation

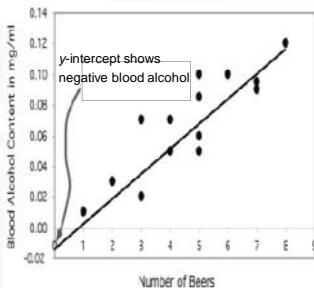
Recall

The y-intercept

Sometimes the y-intercept is **not realistic / not possible**. Here we have negative blood alcohol content, which makes no sense...

But the negative value is appropriate for the equation of the regression line.

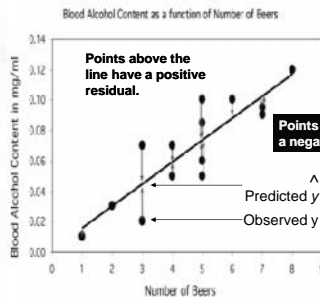
There is a lot of scatter in the data and the line is just an estimate.



Recall

Residuals

The distances from each point to the least-squares regression line give us potentially useful information about the contribution of individual data points to the overall pattern of scatter.



These distances are called "residuals."

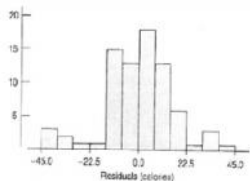
The sum of these residuals is always 0.

residual, $e = y - \hat{y}$

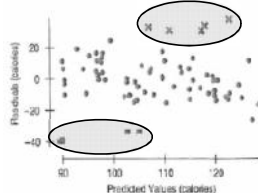
What else can residuals tell us?

- Because a linear regression model is not always appropriate for the data, you should assess the appropriateness of the model by **defining residuals and examining residual plots**. Histograms (and other graphs) of residuals can reveal "Subsets" of data that will enhance our understanding of the original data.
- May lead us to analyzing the "subsets" separately.

Histogram of residuals



Scatterplot of residuals



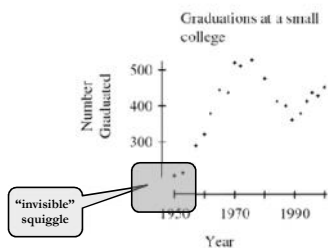
Scatterplots can be deceiving...

- When working with a linear model, we must have data that is linear.
- ALWAYS check the **residual plot** for a pattern when you are trying to verify linearity!

Scatterplots can be deceiving...

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

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

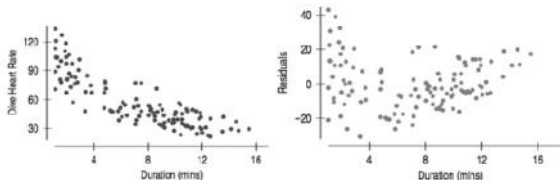


Year	Number Graduated
1950	203
1953	211
1957	288
1960	319
1962	281
1965	446
1968	439
1970	471
1972	509
1976	527
1980	476
1984	413
1987	399
1989	362
1992	379
1994	413
1996	437
1998	426
2000	451

Read the “Penguin Case” and analyze the scatterplots, p. 201-202

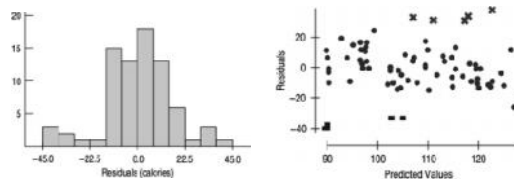
Getting the “Bends”

- 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 Residual from Groups

- 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:



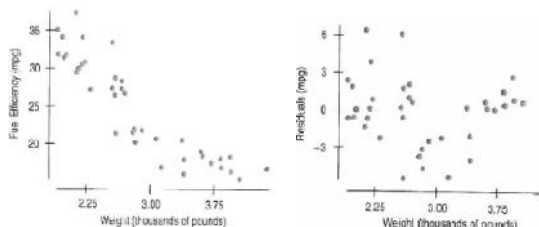
- The small modes in the histogram are marked with different colors and symbols in the residual plot above. What do you see?

Sifting Residual from Groups

- 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.

Hard to see curves...

- Sometimes the scatterplot looks “straight enough”, but a non-linear relationship only comes to light after you look at the **residual plot**.



Getting the “Bends”

- No regression analysis is complete **without a display of the residuals** to check that the linear model is reasonable.
- Because the residuals are what is “left over” after the model describes the relationship, they 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 regression conditions that require our attention.

Getting the “Bends”

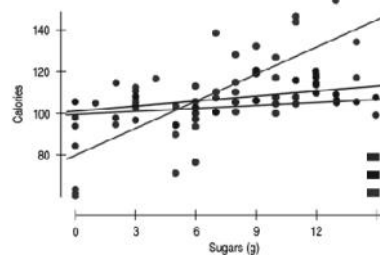
- Linear regression only works for **linear models**. (That sounds obvious, but when you fit a regression, you can’t take it for granted.)
- A curved relationship between two variables might not be apparent when looking at a scatterplot alone, but will be **more obvious in a plot of the residuals**.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

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 necessarily correct. You must determine what makes the most sense. In the following example, we see that modeling them apart makes sense.

Subsets

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

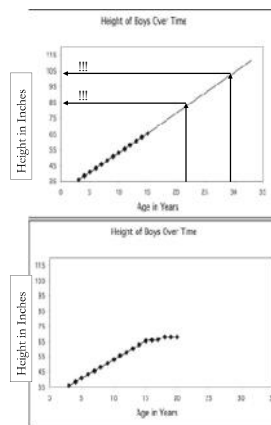


Caution: Extrapolation Ahead!

- When we make predictions with a regression line that are far outside the range of our explanatory (x) values, we are using **extrapolation**. In other words, **extrapolation is the use of a regression line for prediction outside a known range of x values**.
 - ☛ **Extrapolations are often not accurate**
- Associations for variables can be trusted only for the range of values for which data have been collected (known as interpolation)
 - ☛ **Note: even a very strong relationship may not hold outside the data’s range**

Caution: Beware of Extrapolation!

Do the graphs make sense?



Blood Alcohol Content as a function of Number of Beers

There is quite some variation in BAC for the same number of beers drunk. A person's blood volume is a factor in the equation that we have overlooked.

Now we change the number of beers to the number of beers/weight of a person in pounds.

Note how much smaller the variation is. An individual's weight was indeed influencing our response variable "blood alcohol content."

Bacterial growth rate changes over time in closed cultures:

If you only observed bacterial growth in test tubes during a small subset of the time shown here, you could get almost any regression line imaginable. Extrapolation = big mistake

Caution: Beware of Extrapolation!

- Sarah's height was plotted against her age
- Can you predict her height at age 42 months?
- Can you predict her height at age 30 years (360 months)?

Age (months)	36	48	51	54	57	60
Height (cm)	86	90	91	93	94	95

Caution: Beware of Extrapolation!

- Regression line: $y = 71.95 + .383x$
- height at age 42 months? $y = 88$ cm.
- height at age 30 years? $y = 209.8$ cm.
- She is predicted to be 6' 10.5" at age 30!

Making predictions: Interpolation

The equation of the least-squares regression allows you to predict y for any x **within the range studied**. This is called **interpolating**.

Nobody in the study drank 6.5 beers, but by finding the value of \hat{y} from the regression line for $x = 6.5$, we would expect a blood alcohol content of **0.094 mg/ml**.

$\hat{y} = 0.0144 * 6.5 + 0.0008$
 $\hat{y} = 0.936 + 0.0008 = 0.0944 \text{ mg/ml}$

Predicting the Future: Extrapolation

- Extrapolation** is the use of a regression line for predictions outside the range of x -values used to obtain the line.
- Extrapolations can get us in trouble.
- When the x -variable is **Time**, extrapolation becomes an attempt to peer into the future. People have always wanted to see into the future...

Predicting the Future: Extrapolation

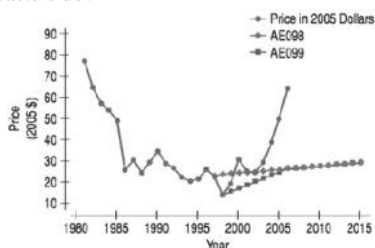
- The model should only be trusted for the span of x -values it represents.
- Extrapolations assume that past trends will continue into the far future.
- An example of extrapolation in the news...
Women may run faster than men in 2156

Extrapolation

- The farther our x -value is from the mean of x , the less we trust our predicted value.
- Once we venture into new x territory, our predicted value is 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** .
- 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.**

Extrapolation

- 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.

Extrapolating from current trends is so tempting that even professional forecasters make this mistake, and sometimes the errors are striking. In the mid-1970s, oil prices surged and long lines at gas stations were common. In 1970, oil cost about \$17 a barrel (in 2005 dollars)—about what it had cost for 20 years or so. But then, within just a few years, the price surged to over \$40. In 1975, a survey of 15 top econometric forecasting models (built by groups that included Nobel prize-winning economists) found predictions for 1985 oil prices that ranged from \$300 to over \$700 a barrel (in 2005 dollars). How close were these forecasts?

FIGURE 9.6
 The scatterplot shows an average increase in the price of a barrel of oil of over \$7 per year from 1971 to 1982.

When the Data Are Years, ... we usually don't enter them as four-digit numbers. Here we used 0 for 1970, 10 for 1980, and so on. Or, we may simply enter two digits, using 82 for 1982, for instance. Rescaling years like this often makes calculations easier and equations simpler. We recommend you do it, too. But be careful! If 1982 is 82, then 2014 is 114.

Outliers, Leverage and Influence

Outliers

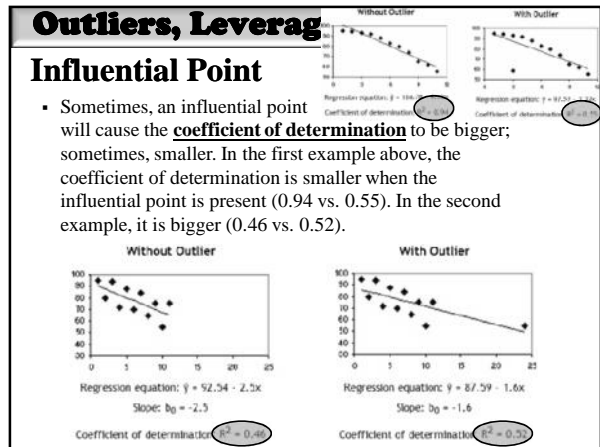
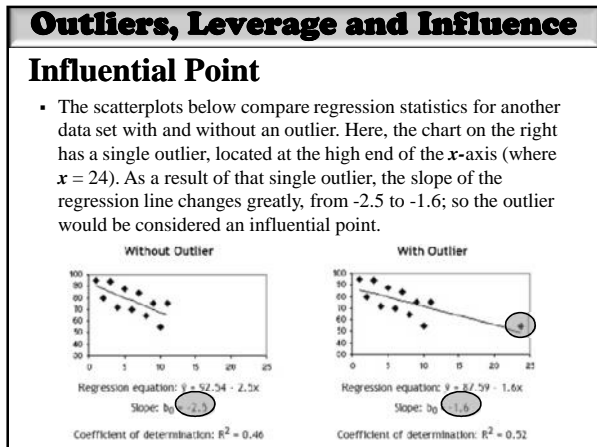
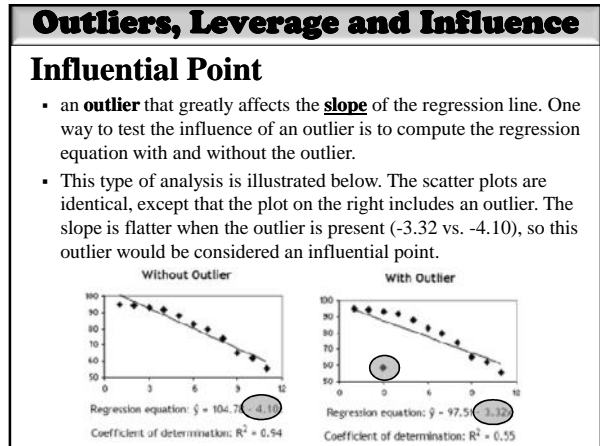
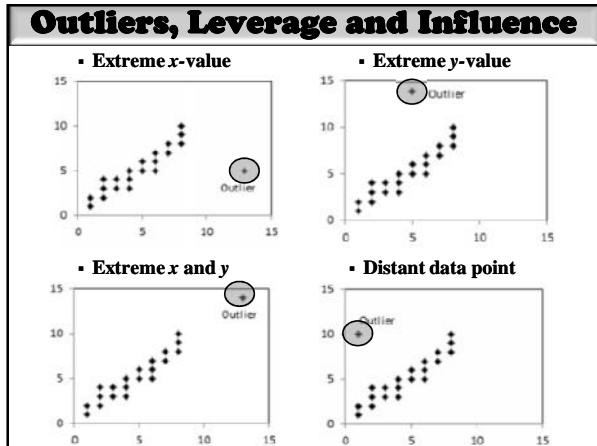
- Data points that stand away from the others/ diverge in a big way from the overall pattern .
- Outlying points can strongly influence a regression. Even a single point far from the body of the data can dominate the analysis.
- Outliers can be extraordinary by having **large residuals** or by having **high leverage**.

Outliers, Leverage and Influence

Outliers

There are four ways that a data point might be considered an outlier.

- It could have an extreme x -value compared to other data points.
- It could have an extreme y -value compared to other data points.
- It could have extreme x and y values.
- It might be distant from the rest of the data, even without extreme x or y values.



Outliers, Leverage and Influence

Influential Point

If your data set includes an influential point, here are some things to consider:

- An influential point may represent bad data, possibly the result of measurement error. If possible, check the validity of the data point.
- Compare the decisions that would be made based on regression equations defined with and without the influential point. If the equations lead to contrary decisions, use caution.

Outliers, Leverage and Influence

Test your understanding of this lesson...

In the context of regression analysis, which of the following statements are true?

- When the data set includes an influential point, the data set is nonlinear.
- Influential points always reduce the coefficient of determination.
- All outliers are influential data points.

(A) I only
 (B) II only
 (C) III only
 (D) All of the above
 (E) None of the above

Data sets with influential points can be linear or nonlinear. In this lesson, we went over an example in which an **influential point** increased the coefficient of determination. With respect to regression, **outliers** are influential only if they have a big effect on the regression equation. Sometimes, outliers do not have big effects. For example, when the data set is very large, a single outlier may not have a big effect on the regression equation.

Outliers, Leverage and Influence

Influential Point

- A point is **influential** if omitting it from the analysis gives a very different model (changes the slope of the line)
- High leverage points can also be influential, but do not need to be
- Not all outliers and leverage points are influential
 - Fit the regression line with and without to determine the influence

Outliers, Leverage and Influence

Influential Point

- A leverage point that does goes against the overall pattern of the data is an **influential point**.
- However, points that are at the edges of our data's range of x -values that go against the overall pattern will also be influential.
- Outliers whose x -value lies in the center of the range of x -values will NOT be influential; they do not affect the linear model much. However, such points WILL weaken the correlation (r).

Outliers, Leverage and Influence

Influential Point

- **Influential points** are points with **high leverage**. They highly influence the slope of the regression line and the correlation coefficient.
- Influential points can be more easily seen in scatterplots of the original data or by finding a regression model with and without the points.
- The surest way to verify that a point is **influential** is to find the regression line with and without the suspect point. **If the line moves more than a small amount when the point is deleted, the point is influential (for the LSRL).**

Outliers, Leverage and Influence

Leverage

- Data points whose x -values are far from the mean of x are said to exert leverage on a linear model.
- Points that are extraordinary in their x -value can especially influence a regression model. We say that they have **high leverage**.
- Can have large effect on the line—**high leverage points pull the regression line close to them**, sometimes completely determining the slope and intercept.
- With high enough leverage, **their residuals can appear deceptively small**. (p. 205)

Outliers, Leverage and Influence

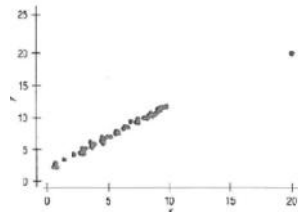
Leverage

- **Leverage points** can either confirm the pattern of data, or they may go against the pattern.
- If they confirm the pattern, the $|r|$ value is increased by the leverage point's presence.
- In other words, if the range of x -values for which a pattern applies is widened/extended, the correlation will increase.
- Imagine a fulcrum (for a see-saw) in the scatterplot, located at the central x -values... a leverage point is a point far out on the see-saw.
- Leverage points can have a big effect on r or on our linear model... Just as a small person can have a big effect on a see-saw if they are seated far enough from the fulcrum.

Types of Unusual Points

1) High Leverage points with small residuals

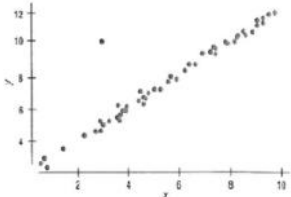
- These points confirm the pattern, but are extreme values. The slope and intercept are mostly unaffected, but the r^2 value will increase—don't be misled that the model is now stronger.



Types of Unusual Points

2) Outliers—Not high leverage, not influential and large residual

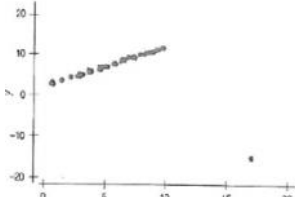
- Does not affect the slope, but aren't consistent with pattern. Will change the intercept. Don't throw away. x value is near center of mean of x -values.



Types of Unusual Points

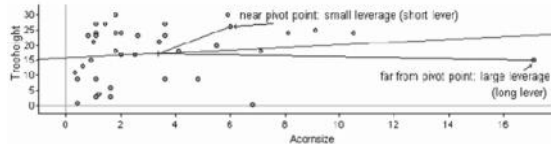
3) Influential Points—also high leverage and probably residual

- These are most troublesome. They aren't consistent with model and if the point is removed the slope of line dramatically changes—it changes the model. Don't throw it out without thinking.



Types of Unusual Points

- Typically, a point that is an outlier in the x -direction will exert influence on the line. "Points tug at the regression line in an attempt to make their residuals smaller. But the regression line pivots around the mean-mean point. Points close to that fulcrum (left to right) can't make their residuals much smaller, hence they do not change the slope of the line much. Points far away (in the x -direction) can exert a lot of leverage—changes in the slope can make their residuals much smaller."



Beware of outliers. You can't interpret a correlation coefficient safely without a background check for outliers. Here's a silly example:

The relationship between IQ and shoe size among comedians shows a surprisingly strong positive correlation of 0.50. To check assumptions, we look at the scatterplot:

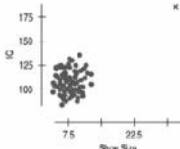



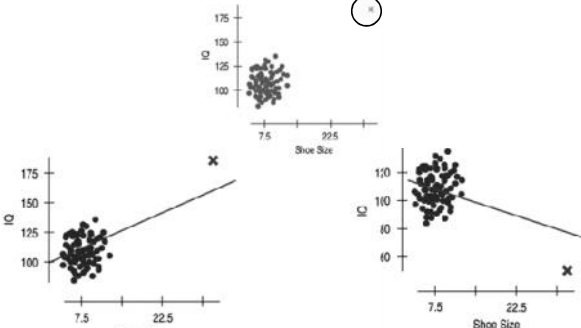
FIGURE 710
A scatterplot of IQ vs. Shoe Size. From this "study," what is the relationship between the two? The correlation is 0.50. Who does that point (the green x) in the upper right-hand corner belong to?



The outlier is Bozo the Clown, known for his large shoes, and widely acknowledged to be a comic "genius." Without Bozo, the correlation is near zero. Even a single outlier can dominate the correlation value. That's why you need to check the Outlier Condition.

Outliers, Leverage and Influence

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



Outliers, Leverage and Influence

- 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

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.
- When a point with **high leverage** lines up with the rest of the data, it **doesn't influence the slope** but it does increase the R^2 .
- Removing a point that is an **x-outlier (high leverage)** but not a model outlier can actually decrease your R^2 .

JUST CHECKING

Each of these scatterplots shows an unusual point. For each, tell whether the point is a high-leverage point, would have a large residual, or is influential.

- 1) Not high leverage, not influential, large residual
- 2) High leverage, not influential, small residual
- 3) High leverage, influential, not large residual

Beware the Lurking Variable!

A **lurking variable**: a variable that is not included in the study but still potentially affects the relationship among the variables in a study

- A **lurking variable can falsely suggest a strong relationship between x and y or it can hide a relationship that is really there**
- Sometimes the relationship between two variables is influenced by other variables that we did not measure or even think about

Lurking Variables and Causation

- With observational data, as opposed to designed experiments, there is not way to be sure that a **lurking variable** is not the cause of any apparent association.
- The lurking variable is some third variable (not the explanatory or predictor variable) that is driving both variables you have observed.
- A **lurking variable** is sometimes referred to as **common response**. It's a variable that drives two other variables, creating the impression of an association between them. For countries, pick any measure of technological modernity (# of TVs per capita) and life expectancy. You'll clearly see an association—countries with fewer TVs have lower life expectancy. Such lurking variables as general economic well-being and standard of living probably explain both. We don't think that having a TV increases your lifespan.

Beware the Lurking Variable!



- There's this guy who's going to clean the windows of a mental asylum. A patient follows him shouts to him "I gotta secret, I gotta secret...", he ignores the patient. Again the patient follows him, but he ignores his cries. By the time he's nearly finished the building, he's really curious about what the patients secret is, so he decides to ask the patient. The patient pulls a matchbox out of his pocket, opens it and puts it on a table. Out crawls this little spider. The patient says "spider go left," and the spider walks to it's left a bit. Then he says "spider go right," the spider walks to its right a little bit.

Beware the Lurking Variable!

- He says "spider turn around, walk forward then go right," and sure enough the spider turns around, walks forward, and then goes right a bit. The window cleaner is amazed "Wow! He says, that's amazing!" "No, that's not my secret," says the patient, "watch." He picks up the spider in his hand and pulls all its legs off then puts it back on the table. "Spider go right," the spider doesn't move, "spider go left," the spider doesn't move, "Spider, turn around" again the spider doesn't move. "There!" he says, "that's my secret, if you pull all the spider's legs off they go deaf....."

Lurking variables

- Describe the association.
- What is the lurking variable in these examples?
- How could you answer if you didn't know anything about the topic?

• Strong positive association between the number of firefighters at a fire site and the amount of damage a fire does



the number of firefighters at a fire site vs. the amount of damage a fire does

• Negative association between moderate amounts of wine drinking and death rates from heart disease in developed nations



amount of wine drinking vs. death rates from heart disease in developed nations

How to spot the presence of the lurking variable?

- Because lurking variables are often unrecognized and unmeasured, detecting their effect is a challenge.
- Many lurking variables change systematically over time.
 - Plot both the response variable and the regression residuals against the time order of the observations whenever possible. An understanding of the background of the data then allows you to guess what lurking variables might be present.

Example: Discrimination in Medical Treatment?

- Studies show that men who complain of chest pain are more likely to get detailed tests and aggressive treatment such as bypass surgery than are women with similar complaints. Is this association between gender and treatment due to discrimination?
 - Not necessarily. Men and women develop heart problems at different ages –women are on the average between 10 and 15 years older than men. Aggressive treatments are more risky for older patients, so doctor's may hesitate to recommend them. Lurking variables—the patients age and condition—may explain the relationship between gender and doctors' decisions.

Example: TV and Life Expectancy

- Measure the number of television sets per person x and the average life expectancy y for the world's nations. There is a high correlation: nations with many TV sets have higher life expectancies.
- Could we lengthen the lives of people in Rwanda by shipping them TV sets ?



"In a new attack on third world poverty, aid organizations today began delivery of 100,000 television sets."

Example: TV and Life Expectancy

No. Rich nations have more TV sets than poor nations. Rich nations also have longer life expectancies because they offer better nutrition, clean water, and better health care. Clearly, there is no cause and effect relationship between TV sets and length of life.

A Last Lurking Variable Example

- A study showed that there was a strong correlation between the number of firefighters at a fire and the property damage that the fire causes.
 - So maybe we should send less firefighters to fight fires?
 - WRONG! If the fire is severe and/or already large, we should send more firefighters to fight the fire.



🍷* **Causation vs. Association** 🍷*

- Some studies want to find the existence of **causation**.

Examples of causation:

- 👉 Increased drinking of alcohol causes a decrease in coordination.
- 👉 Smoking and Lung Cancer.



Examples of association:

- 👉 High SAT scores are associated with a high Freshman year GPA.
- 👉 Smoking and Lung Cancer.

Correlation Does NOT Imply Causation !



Even very strong correlations may not correspond to a real causal relationship.



🍷* **Evidence of Causation** 🍷*

- A properly conducted experiment establishes the connection
- Other considerations:
 - ✓ A reasonable explanation for a cause and effect exists
 - ✓ The connection happens in repeated trials
 - ✓ The connection happens under varying conditions
 - ✓ Potential confounding factors are ruled out
 - ✓ Alleged cause precedes the effect in time

Reasons Two Variables May Be Related (Correlated)

- Explanatory variable causes change in response variable
- Response variable causes change in explanatory variable
- Explanatory may have some cause, but is not the sole cause of changes in the response variable
- Confounding variables may exist
- Both variables may result from a common cause such as, both variables *changing over time*
- The correlation may be merely a coincidence

Explanatory Causes Response

- **Explanatory:** pollen count from grasses
- **Response:** percentage of people suffering from allergy symptoms
- **Explanatory:** amount of food eaten
- **Response:** hunger level

Response Causes Explanatory

- **Explanatory:** Hotel advertising dollars
- **Response:** percentage Occupancy rate
- **Positive correlation?** – more advertising leads to increased occupancy rate?
 - 🍷 **Actual correlation is negative:** lower occupancy leads to more advertising

Explanatory is NOT Sole Contributor

- **Explanatory:** Consumption of barbecued foods
- **Response:** percentage Incidence of stomach cancer
- *Barbecued foods are known to contain carcinogens, but other lifestyle choices may also contribute*

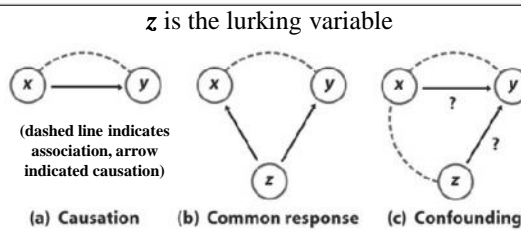
*** Association does not imply Causation ***

- An association between two variables x and y can reflect many types of relationship among x , y , and one or more lurking variables.
- An association between an explanatory variable (predictor) x and a response variable y , even if it is very strong, is not by itself good evidence that changes in x actually cause changes in y .

*** How to show Causation? ***

- The only way to get absolutely conclusive evidence of cause and effect or that x causes changes in y is to do an *experiment* in which we change x in an environment which we completely control, this keeps lurking variables under control
- When experiments cannot be done, finding the explanation for an observed association is often difficult and can even be controversial

Lurking Variables and Causation



- Even a very strong association between two variables is **NOT** by itself a good evidence that there is a cause-and-effect link between the variables.
- The main question of establishing causation: How can a direct causal link between x and y be established?

Explaining Association: Direct Causation

• **Cause-and-effect**

Examples:

- ⇒ Amount of fertilizer and yield of corn
- ⇒ Weight of a car and its MPG
- ⇒ Dosage of a drug and the survival rate of the mice



Causation
(a)

• We already know x and y are associated. The arrow shows causation between the two variables, i.e., “ x causes y .”

Explaining Association: Direct Causation

x = mom’s adult height
 y = daughter’s adult height

• Experiments have shown that the mom’s adult height is an appropriate predictor for a daughter’s adult height. We have association AND causation.

x = mother’s body mass index (BMI)
 y = daughter’s body mass index (BMI)

- Body part is determined by heredity. (based on a study). Daughters inherit half their genes from their mothers. There is therefore a direct causal link between the BMI of mothers and daughters.
- CAUTION: Even when direct causation is present, it is rarely a complete explanation of an association between two variables.

Lurking Variables and Causation

Causation

- Does smoking cause cancer?
- Did chemical weapons exposure cause health problems in Gulf War vets?
- Will increasing the speed limit increase traffic fatalities?
- Will bringing storks into an area increase the birth rate?
- Will lowering the drinking age limit in California affect the university dorm drinking parties?
- High temperatures in the summer lead to higher electricity use (fans, air conditioning, etc)

Lurking Variables and Causation

Causation

Example:

- Brothers and sisters heights are highly correlated. However a tall brother doesn't cause a tall sister.
- A more likely cause: common genetics
- Even though there may no be a causal relationship between two variables, it can still be useful to predict a sister's height from her brother's height.

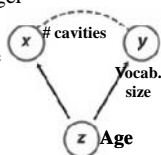
Lurking Variables and Causation

Common Response

- Refers to the possibility that a change in a lurking variable is causing changes in both our explanatory variable and our response variable.

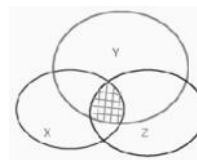
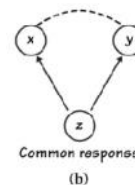
Example:

- It has been observed that children with more cavities tend to have larger vocabularies. However it is hard to see how more cavities might lead to larger vocabularies (or vice versa).
- However in this case, both variables are associated with **age**.



Explaining Association: Common Response

- Both x and y change in response to changes in z , the lurking variable
- There may not be direct causal link between x and y .
- The lurking variable distort the true relation between x and y .
- Lurking variables can create nonsense correlations!



⚡ *x and y show an association but it is really the lurking variable z doing the work.*

Explaining Association: Common Response

x = a high school senior's SAT score
 y = the student's first-year college grade point average


- *Students who are smart and who have learned a lot tend to have both high SAT scores and high college grades. The positive correlation is explained by this common response (lurking variable) to students' ability and knowledge. Bright students would tend to do well on both.*

Explaining Association: Common Response

x = monthly flow of money into stock mutual funds
 y = monthly rate of return for the stock market

- *There is a strong positive correlation between how much money individuals add to mutual funds each month and how well the stock market does the same month.*
- *Is the new money driving the market up?*
- *The correlation may be explained in part by common response to underlying investor sentiment: when optimism reigns, individuals send money to funds and large institutions also invest more.*
- *The institutions would drive up prices even if individuals did nothing. In addition, what causation there is may operate in the other direction: when the market is doing well, individuals rush to add money to their mutual funds.*

Explaining Association: Common Response



x = Divorce among men
 y = Percent abusing alcohol

- Both variables change due to common cause
- Both may result from an unhappy marriage.

Explaining Association: Common Response

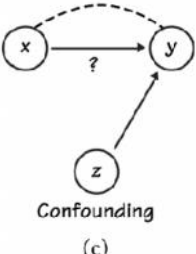
• Both Variables are Changing Over Time

- Both divorces and suicides have increased dramatically since 1900.
- Are divorces causing suicides?
- Are suicides causing divorces???
- The population has increased dramatically since 1900 (causing both to increase).
 - Better to investigate: Has the rate of divorce or the rate of suicide changed over time?

Explaining Association: Confounding

- Two variables (whether explanatory or lurking) are **confounded** when their effects on a response variable cannot be distinguished from each other.

• Again, x and y show an association, but in this case, we are unable to determine whether x is causing y or if z is causing y .



Confounding
(c)

Lurking Variables and Causation

Confounding

- Two variables are **confounded** when you can't tell which of them (or whether it's the combination) had an affect.
- Refers to the possibility that either the change in our explanatory variable is causing changes in the response variable OR that a change in a lurking variable is causing changes in the response variable.
- You might want to test a fertilizer on your lawn. Suppose you spread it on half the lawn to see if the grass will look better there. If you spread it on the sunny half, leaving the shady half unfertilized, you won't know whether the greener grass resulted from fertilizer or sunshine (or the two together).

Explaining Association: Confounding

x = whether a person regularly attends religious services
 y = how long the person lives

- Many studies have found that people who are active in their religion live longer than nonreligious people.
- But people who attend church or mosque or synagogue also take better care of themselves than non-attenders. They are less likely to smoke, more likely to exercise, and less likely to be overweight.
- The effects of these good habits are confounded with the direct effects of attending religious services.

Explaining Association: Confounding

x = Meditation
 y = Aging (measurable aging factor)

- General concern for one's well being may be confounded with decision to try meditation

Explaining Association: Confounding

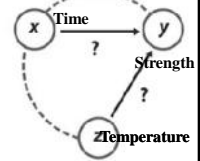
x = the number of years of education a worker has
 y = the worker's income

- It is likely that more education is a cause of higher income—many highly paid professions require advanced education.
- However, confounding is also present. People who have high ability and come from prosperous homes are more likely to get many years of education than people who are less able or poorer.
- Of course, people who start out able and rich are more likely to have high earnings even without much education. We can't say how much of the higher income of well-educated people is actually caused by their education.

Lurking Variables and Causation

Confounding

Example: Strength of molded parts
 x : time in mold y : strength of part



- In a study, higher strength was associated with longer mold times.
- The way the experiment was performed was to have all the samples at 10 seconds in the mold done first, then the samples at 20 seconds, then 30 seconds, and so on.
- They also saw a strong relationship between strength and the order done. The time in the mold and the order done were **confounded**.
- It ended up that the mold got **warmer** as more batches were done and higher temperature increases strength.

Lurking Variables and Causation

Confounding

- Suppose you want to compare laundry detergent A vs detergent B. You wash a bunch of loads using A and B. But you always put A in washer #1 and always put B in #2. Now you're confounded. You don't know if it's the detergent or the washing machine that made one load cleaner than the other.
- A store's special promotion may increase video rentals but the marketing folks cannot be sure that's what did it if the weather was particularly bad during the trial period. Bad weather may have kept people indoors and induced them to rent more videos anyway. Any actual effect of the special promotion is confounded by the weather.

Lurking Variables and Causation

Common response

- When more kids eat ice cream, more kids drown. It's the warmer weather that's causing an increase in both. They are both responded to summer.
- Student who are smart and who have learned a lot tend to have both high SAT scores and high college grades. The positive correlation is explained by the common response to students' ability and knowledge.



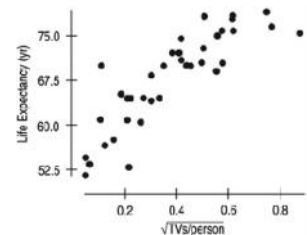
Lurking Variables and Causation

- There have been many studies showing a strong positive association between hours spent in religious activities (going to church, attending religious classes, praying, etc) and life expectancy. NOT CAUSATION. There is confounding—on average, people who attend religious activities also take better care of themselves than non-church attendants. They are also less likely to smoke, more likely to exercise and less likely to be overweight. These effects of **good habits** (lurking variables) are confounded with the direct effects of attending religious activities.

Lurking Variables and Causation

- Measure the number of television sets per person x and the average life expectancy y for the world's nations.
- There is a high positive correlation: nations with many TV sets have higher life expectancies. Could we lengthen the lives of people in Rwanda by shipping them TV sets?

- The 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

- 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.
- The variables x and y have a common response variable z , per capita income. While, nations with higher per capita income have more TV sets than do poor nations, they also tend to have better nutrition, cleaner water, and better health care.

Lurking Variables and Causation

Does smoking cause lung cancer?



- In order to know that smoking causes cancer, we would have to design an experiment where we can change the explanatory variable (smoking or not) in a controlled environment.
 - ✓ *Can we ethically make people smoke, drink, do illicit drugs, etc.?*
 - ✓ *Are there other types of cause and effect relationships similar to this scenario?*

Lurking Variables and Causation

Does smoking cause lung cancer?



- **Proving smoking causes lung cancer**
 - ✓ *Association between smoking and lung cancer is very strong*
 - ✓ *This association is consistent in many studies*
 - ⊕ *Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.*
 - ✓ *High doses are associated with stronger responses*
 - ⊕ *That is, people who smoke more often tend to get lung cancer more often*

Lurking Variables and Causation

Does smoking cause lung cancer?



- **Proving smoking causes lung cancer**
 - ✓ *The alleged cause precedes the effect in time*
 - ⊕ *Lung cancer develops after years of smoking. It kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke.*
 - ✓ *The alleged cause is plausible*
 - ⊕ *Experiments with animals show that tars from cigarette smoke do cause cancer.*
 - ✓ *Still, in some cases, a person might get lung cancer from pollution, working in a factory, etc.*

Lurking Variables and Causation

Does smoking cause lung cancer?



- **Causation:** smoking causes lung cancer.
- **Common response:** people who have a genetic predisposition to lung cancer also have a genetic predisposition to smoking.
- **Confounding:** people who drink too much, don't exercise, eat unhealthy foods, etc. are more likely to get lung cancer as a result of their lifestyle. Such people may be more likely to be smokers as well.

Lurking Variables and Causation

Car weight and gas mileage

- **Causation:** Physics says the more weight, the more energy you need to move an object, therefore implying worse gas mileage.
- **Common response:** The type of car (van, sports, SUV, etc) influences the weight, plus other factors that affect the gas mileage of the car.
- **Confounding:** While weight has a causative effect, its actual effect can not be accurately ascertained since weight is confounded with a number of factors, such as engine size or horsepower.

- Some other examples
- Rainfall amounts and plant growth
 - Explanatory variable – rainfall
 - Response variable – plant growth
 - Possible lurking variable – amount of sunlight
- Exercise and cholesterol levels
 - Explanatory variable – amount of exercise
 - Response variable – cholesterol level
 - Possible lurking variable – diet

Examples

- Children who watch many hours of TV get lower grades in school on average than those who watch less TV
 - Does this mean that TV *causes* poor grades?
 - What are potential confounding variables?
- People who use artificial sweeteners in place of sugar tend to be heavier than people who use sugar
 - Does this mean that sweeteners *cause* weight gain?
 - What is probably happening here?

More examples of Causation errors...

- a) The number of firefighters at a fire vs. the amount of damage in dollars
- b) The number of hours students work vs. their grades
- c) The number of hours of TV watched by young children and the length of their attention span

Establishing Causation

- **The only compelling method: Designed experiment (Chapters 12-13)**
- **Hot disputes:**
 - ↳ *Does gun control reduce violent crime?*
 - ↳ *Does living near power lines cause cancer?*
 - ↳ *Does smoking cause lung cancer?*
 - ↳ *Does the use of fossil fuel causing global warming?*

Beware of Correlations based on Averaged Data!

- **Many regression and correlation studies work with averages or other measures that combine information from many individuals**
 - ☞ *Note when researchers use such techniques. Resist the temptation to apply the results of such studies to individuals.*
- **Correlations based on averages are usually too high when applied to individuals.**
 - ☞ *Note exactly what variables were measured in a statistical study.*

Working with Summary Values

- Be cautious when working with data values that are summaries, such as mean and medians.
- These values have less variability and therefore inflate the strength of the relationship (correlation).

Working With Summary Values

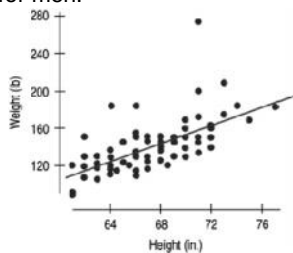
- Scatterplots of statistics summarized over groups tend to show less variability than we would see if we measured the same variable on individuals.
- This is because the summary statistics themselves vary less than the data on the individuals do.

Working With Summary Values

- Scatterplots of statistics summarized over groups tend to show LESS variability than we would see if we measured the same variable on individuals.
- For example, consider plotting the height and weights of students in a certain grade.
- Then imagine that instead of plotting a point for each individual student, we find the average weight for each height and plot the heights vs. the average weights.
- This second scatterplot is likely to have a lot less scatter and a higher R^2 value.
- Why? Because means vary less than individual values do.
- Sort of similar to what we saw with Simpson's paradox; sometimes lumping data together (in this case by using summary stats) we LOSE information.

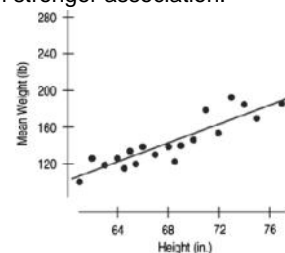
Working With Summary Values (cont.)

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



Working With Summary Values (cont.)

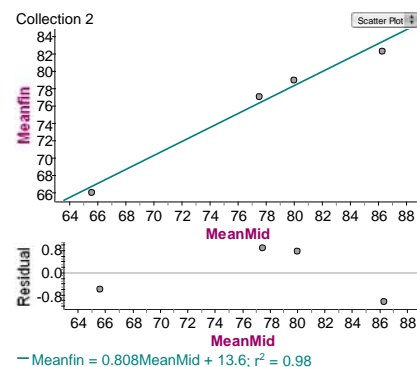
- If instead of data on individuals we only had the mean weight for each height value, we would see an even stronger association:

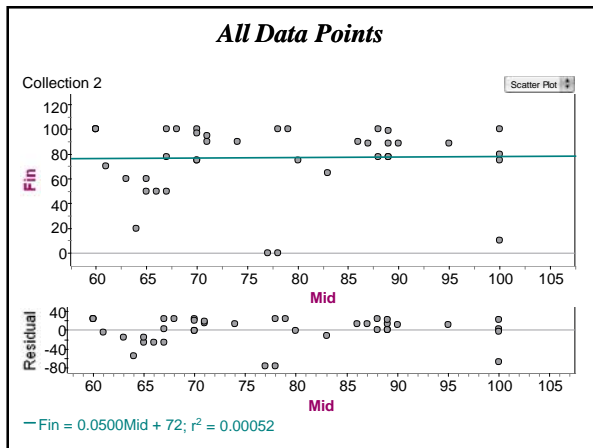


Working With Summary Values (cont.)

- 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.

Summary Data





When you make a linear model:

- Make sure the relationship is straight.
- Beware of extrapolating.
- Beware of especially extrapolating into the future.
- Be on guard for different groups in your regression.
- Look for outliers.
- Beware of high leverage points, and especially of those that are influential.
- Consider comparing two regressions.
- Treat outliers honestly.
- Beware of lurking variables.
- Watch out when dealing with data that are summaries.