STA 2023

Module 5
Regression and Correlation

Learning Objectives

Upon completing this module, you should be able to:

1. Define and apply the concepts related to linear equations with one independent variable.
2. Explain the least-squares criterion.
3. Obtain and graph the regression equation for a set of data points, interpret the slope of the regression line, and use the regression equation to make predictions.
4. Define and use the terminology predictor variable and response variable.
5. Understand the concept of extrapolation.
6. Identify outliers and influential observations.
7. Know when obtaining a regression line for a set of data points is appropriate.

2

Learning Objectives (Cont.)

8. Determine and interpret the coefficient of determination.
9. Determine and interpret the linear correlation coefficient, *r*.
10. Explain and apply the relationship between the linear correlation coefficient and the coefficient of determination.

3

What are Scatterplots?

Scatterplots may be the most common and most effective display for data.

- In a scatterplot, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others.

Scatterplots are the best way to start observing the relationship and the ideal way to picture associations between two *quantitative* variables.

4

Looking at Scatterplots

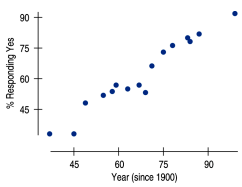
When looking at scatterplots, we will look for direction, form, strength, and unusual features.

Direction:

- A pattern that runs from the upper left to the lower right is said to have a negative direction.
- A trend running the other way has a positive direction.

5

Looking at Scatterplots (cont.)

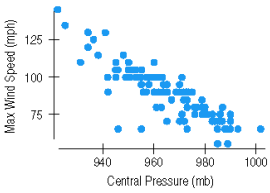


This figure shows a positive association between the year since 1900 and the % of people who say they would vote for a woman president.

As the years have passed, the percentage who would vote for a woman has increased.

6

Looking at Scatterplots (cont.)



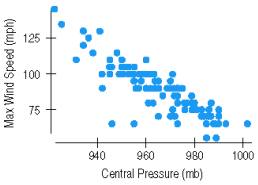
This figure shows a negative association between central pressure and maximum wind speed

As the central pressure increases, the maximum wind speed decreases.

Looking at Scatterplots (cont.)

Form:

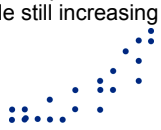
– If there is a straight line (linear) relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form.



Looking at Scatterplots (cont.)

Form:

– If the relationship isn't straight, but curves gently, while still increasing or decreasing steadily,



we can often find ways to make it more nearly straight.

Looking at Scatterplots (cont.)

Form:

– If the relationship curves sharply,

the methods of this module cannot really help us.

10

Looking at Scatterplots (cont.)

Strength:

– At one extreme, the points appear to follow a single stream

(whether straight, curved, or bending all over the place).

11

Looking at Scatterplots (cont.)

Strength:

– At the other extreme, the points appear as a vague cloud with no discernable trend or pattern:

– Note: we will quantify the amount of scatter soon.

12

Looking at Scatterplots (cont.)

Unusual features:

- Look for the unexpected.
- Often the most interesting thing to see in a scatterplot is the thing you never thought to look for.
- One example of such a surprise is an outlier standing away from the overall pattern of the scatterplot.
- Clusters or subgroups should also raise questions.

13

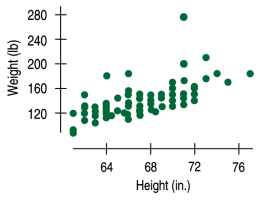
Roles for Variables

- It is important to determine which of the two quantitative variables goes on the *x*-axis and which on the *y*-axis.
- This determination is made based on the roles played by the variables.
- When the roles are clear, the explanatory or predictor variable goes on the *x*-axis, and the response variable goes on the *y*-axis.
- The roles that we choose for variables are more about how we *think* about them rather than about the variables themselves.
- Just placing a variable on the *x*-axis doesn't necessarily mean that it explains or predicts *anything*. And the variable on the *y*-axis may not respond to it in any way.

14

Correlation

Data collected from students in Statistics classes included their heights (in inches) and weights (in pounds):

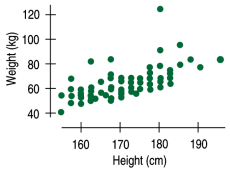


- Here we see a positive association and a fairly straight form, although there seems to be a high outlier.

15

Correlation (cont.)

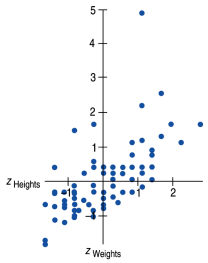
- How strong is the association between weight and height of Statistics students?
- If we had to put a number on the strength, we would not want it to depend on the units we used.
- A scatterplot of heights (in centimeters) and weights (in kilograms) doesn't change the shape of the pattern:



16

Correlation (cont.)

- Since the units don't matter, why not remove them altogether?
- We could standardize both variables and write the coordinates of a point as (z_x, z_y).
- Here is a scatterplot of the standardized weights and heights:



17

Correlation (cont.)

- Note that the underlying linear pattern seems steeper in the standardized plot than in the original scatterplot.
- That's because we made the scales of the axes the same.
- Equal scaling gives a neutral way of drawing the scatterplot and a fairer impression of the strength of the association.

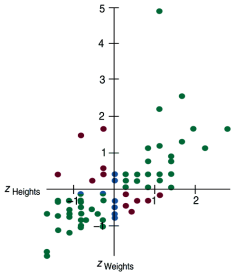
18

Correlation (cont.)

Some points (those in green) strengthen the impression of a positive association between height and weight.

Other points (those in red) tend to weaken the positive association.

Points with *z*-scores of zero (those in blue) don't vote either way.



19

Correlation (cont.)

The correlation coefficient (*r*) gives us a numerical measurement of the strength of the linear relationship between the explanatory and response variables.

For the students' heights and weights, the correlation is 0.644. What does this mean in terms of strength? We'll address this shortly.

20

Correlation Conditions

Correlation measures the strength of the *linear* association between two *quantitative* variables.

Before you use correlation, you must check several conditions:

- Quantitative Variables Condition
- Straight Enough Condition
- Outlier Condition

21

Correlation Conditions (cont.)

Quantitative Variables Condition:

- Correlation applies only to quantitative variables.
- Don't apply correlation to categorical data masquerading as quantitative.
- Check that you know the variables' units and what they measure.

22

Correlation Conditions (cont.)

Straight Enough Condition:

- You can *calculate* a correlation coefficient for any pair of variables.
- But correlation measures the strength only of the *linear* association, and will be misleading if the relationship is not linear.

23

Correlation Conditions (cont.)

Outlier Condition:

- Outliers can distort the correlation dramatically.
- An outlier can make an otherwise small correlation look big or hide a large correlation.
- It can even give an otherwise positive association a negative correlation coefficient (and vice versa).
- When you see an outlier, it's often a good idea to report the correlations with and without the point.

24

Correlation Properties

The sign of a correlation coefficient gives the direction of the association.

Correlation is always between -1 and +1.

- Correlation *can* be exactly equal to -1 or +1, but these values are unusual in real data because they mean that all the data points fall *exactly* on a single straight line.
- A correlation near zero corresponds to a weak linear association.

25

Correlation Properties (cont.)

- Correlation treats *x* and *y* symmetrically:
  - The correlation of *x* with *y* is the same as the correlation of *y* with *x*.
- Correlation has no units.
- Correlation is not affected by changes in the center or scale of either variable.
  - Correlation depends only on the *z*-scores, and they are unaffected by changes in center or scale.

26

Correlation Properties (cont.)

- Correlation measures the strength of the *linear* association between the two variables.
  - Variables can have a strong association but still have a small correlation if the association isn't linear.
- Correlation is sensitive to outliers. A single outlying value can make a small correlation large or make a large one small.

27

Correlation Tables

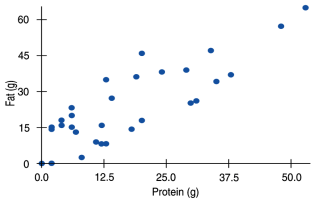
It is common in some fields to compute the correlations between each pair of variables in a collection of variables and arrange these correlations in a table.

	Assets	Sales	Market Value	Profits	Cash Flow	Employees
Assets	1.000					
Sales	0.746	1.000				
Market Value	0.682	0.879	1.000			
Profits	0.602	0.814	0.968	1.000		
Cash Flow	0.641	0.855	0.970	0.989	1.000	
Employees	0.594	0.924	0.818	0.762	0.787	1.000

28

Example

The following is a scatterplot of total *fat* versus *protein* for 30 items on the Burger King menu:



29

The Linear Model

- Correlation says "There seems to be a linear association between these two variables," but it doesn't tell *what that association is*.
- We can say more about the linear relationship between two quantitative variables with a model.
- A model simplifies reality to help us understand underlying patterns and relationships.
- The linear model is just an equation of a straight line through the data.
  - The points in the scatterplot don't all line up, but a straight line can summarize the general pattern.
  - The linear model can help us understand how the values are associated.

30

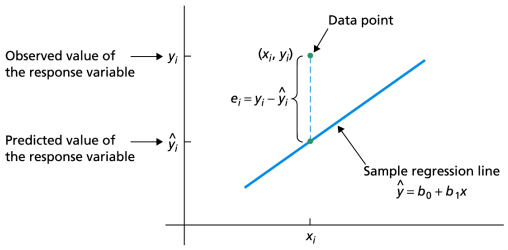
Residuals

- The model won't be perfect, regardless of the line we draw. Some points will be above the line and some will be below. The estimate made from a model is the predicted value, $\hat{y}$.
- The difference between the observed value and its associated predicted value is called the residual.
- To find the residuals, we always subtract the predicted value from the observed one:

$$residual = observed - predicted = y - \hat{y}$$

31

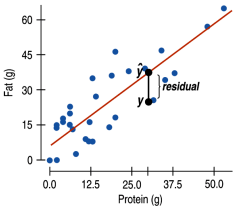
Observed Value and Predicted Value



32

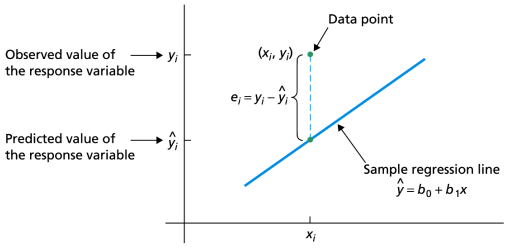
Residuals (cont.)

- A negative residual means the predicted value's too big (an overestimate).
- A positive residual means the predicted value's too small (an underestimate).



33

Observed Value and Predicted Value



34

"Best Fit" Means Least Squares

- Some residuals are positive, others are negative, and, on average, they cancel each other out.
- So, we can't assess how well the line fits by adding up all the residuals.
- Similar to what we did with deviations, we square the residuals and add the squares.
- The smaller the sum, the better the fit.
- The line of best fit is the line for which the sum of the squared residuals is smallest.

35

The Least Squares Line

- We write our model as

$$\hat{y} = b_0 + b_1 x$$

- This model says that our *predictions* from our model follow a straight line.
- If the model is a good one, the data values will scatter closely around it.

36

The Least Squares Line (cont.)

In our model, we have a slope (b_1):

- The slope is built from the correlation and the standard deviations:

$$b_1 = r\frac{s_y}{s_x}$$

- Our slope is always in units of *y* per unit of *x*.

37

The Least Squares Line (cont.)

In our model, we also have an intercept (b_0).

- The intercept is built from the means and the slope:

$$b_0 = \bar{y} - b_1\bar{x}$$

- Our intercept is always in units of *y*.

38

Some Definitions

Least-Squares Criterion

The **least-squares criterion** is that the line that best fits a set of data points is the one having the smallest possible sum of squared errors.

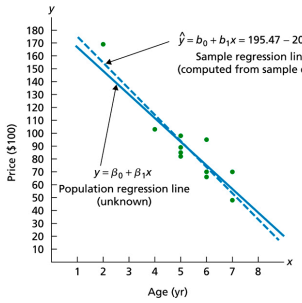
Regression Line and Regression Equation

Regression line: The line that best fits a set of data points according to the least-squares criterion.

Regression equation: The equation of the regression line.

39

Sample Regression Line

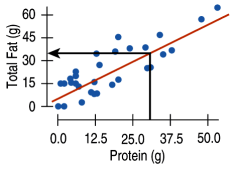


The Regression Line

- The regression line for the Burger King data fits the data well:
  - The equation is

$$\widehat{fat} = 6.8 + 0.97\,protein.$$

  - The *predicted fat* content for a BK Broiler chicken sandwich is
    6.8 + 0.97(30) = 35.9 grams of fat.



The Least Squares Line (cont.)

Since regression and correlation are closely related, we need to check the same conditions for regressions as we did for correlations:

- Quantitative Variables Condition
- Straight Enough Condition
- Outlier Condition

Linear Correlation Coefficient

Linear Correlation Coefficient

For a set of n data points, the **linear correlation coefficient, r,** is defined by

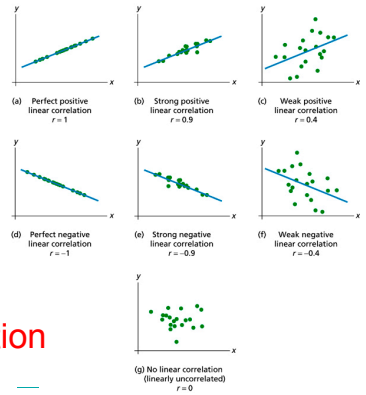
$$r = \frac{\frac{1}{n-1}\Sigma(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$$

where s_x and s_y denote the sample standard deviations of the x-values and y-values, respectively. It can be expressed as $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$. Thus, we can also use the following computing formula to obtain the linear correlation coefficient:

$$r = \frac{\Sigma x_i y_i - (\Sigma x_i)(\Sigma y_i)/n}{\sqrt{\left[\Sigma x_i^2 - (\Sigma x_i)^2/n\right]\left[\Sigma y_i^2 - (\Sigma y_i)^2/n\right]}}.$$

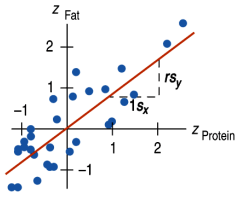
43

Correlation



Correlation and the Line

- Moving one standard deviation away from the mean in *x* moves us *r* standard deviations away from the mean in *y*.
- This relationship is shown in a scatterplot of *z*-scores for *fat* and *protein*.
- Put generally, moving any number of standard deviations away from the mean in *x* moves us *r* times that number of standard deviations away from the mean in *y*.



45

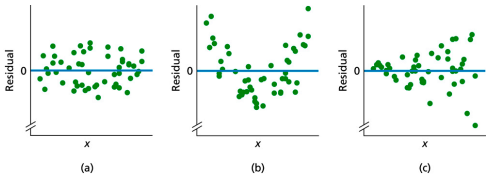
How Big Can Predicted Values Get?

- *r* cannot be bigger than 1 (in absolute value), so each predicted *y* tends to be closer to its mean (in standard deviations) than its corresponding *x* was.
- This property of the linear model is called regression to the mean; the line is called the regression line.

46

Residual Plots

a) no violation of linearity or constant standard deviation
b) Violation of linearity
c) violation of constant standard deviation



47

Residuals Revisited

The linear model assumes that the relationship between the two variables is a perfect straight line. The residuals are the part of the data that *hasn't* been modeled.

Data = Model + Residual

or (equivalently)

Residual = Data – Model

Or, in symbols,

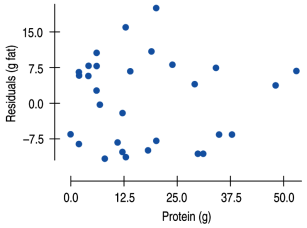
$$e = y - \hat{y}$$

- Residuals help us to see whether the model makes sense.
- When a regression model is appropriate, nothing interesting should be left behind.
- After we fit a regression model, we usually plot the residuals in the hope of finding…nothing.

48

Residuals Revisited (cont.)

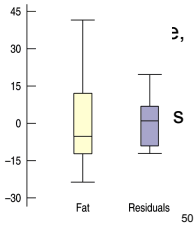
The residuals for the BK menu regression look appropriately boring:



49

R^2—The Variation Accounted For

- The variation in the residuals is the key to assessing how well the model fits.
- In the BK menu items total *fat* has deviation grams. The deviation is 9.2 grams.



50

R^2—The Variation Accounted For (cont.)

- If the correlation were 1.0 and the model predicted the *fat* values perfectly, the residuals would all be zero and have no variation.
- As it is, the correlation is 0.83—not perfection.
- However, we did see that the model residuals had less variation than total *fat* alone.
- We can determine how much of the variation is accounted for by the model and how much is left in the residuals.

51

R^2 - The Variation Accounted For (cont.)

- The squared correlation, r^2, gives the fraction of the data's variance accounted for by the model.
- Thus, $1 - r^2$ is the fraction of the original variance left in the residuals.
- For the BK model, $r^2 = 0.83^2 = 0.69$, so 31% of the variability in total *fat* has been left in the residuals.

52

R^2—The Variation Accounted For (cont.)

- All regression analyses include this statistic, although by tradition, it is written R^2 (pronounced "*R*-squared"). An R^2 of 0 means that none of the variance in the data is in the model; all of it is still in the residuals.
- When interpreting a regression model you need to *Tell* what R^2 means.
  - In the BK example, 69% of the variation in total *fat* is accounted for by the model.

53

How Big Should R^2 Be?

- R^2 is always between 0% and 100%. What makes a "good" R^2 value depends on the kind of data you are analyzing and on what you want to do with it.
- The standard deviation of the residuals can give us more information about the usefulness of the regression by telling us how much scatter there is around the line.
- Along with the slope and intercept for a regression, you should always report R^2 so that readers can judge for themselves how successful the regression is at fitting the data.
- Statistics is about variation, and R^2 measures the success of the regression model in terms of the fraction of the variation of *y* accounted for by the regression.

54

Regression Assumptions and Conditions

- Quantitative Variables Condition:
  - Regression can only be done on two quantitative variables, so make sure to check this condition.
- Straight Enough Condition:
  - The linear model assumes that the relationship between the variables is linear.
  - A scatterplot will let you check that the assumption is reasonable.

55

Regressions Assumptions and Conditions (cont.)

- It's a good idea to check linearity again *after* computing the regression when we can examine the residuals.
- You should also check for outliers, which could change the regression.
- If the data seem to clump or cluster in the scatterplot, that could be a sign of trouble worth looking into further.

56

Regressions Assumptions and Conditions (cont.)

- If the scatterplot is not straight enough, stop here.
  - You can't use a linear model for *any* two variables, even if they are related.
  - They must have a *linear* association or the model won't mean a thing.
- Some nonlinear relationships can be saved by re-expressing the data to make the scatterplot more linear.

57

Regressions Assumptions and Conditions (cont.)

- Outlier Condition:
  - Watch out for outliers.
  - Outlying points can dramatically change a regression model.
  - Outliers can even change the sign of the slope, misleading us about the underlying relationship between the variables.

Is the Regression Reasonable?

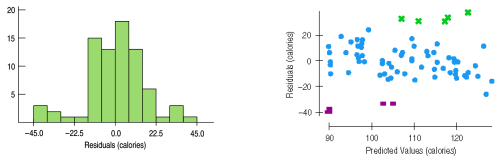
- Statistics don't come out of nowhere. They are based on data.
  - The results of a statistical analysis should reinforce your common sense, not fly in its face.
  - If the results are surprising, then either you've learned something new about the world or your analysis is wrong.
- When you perform a regression, think about the coefficients and ask yourself whether they make sense.

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.

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:



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

61

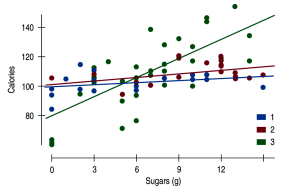
Sifting Residuals for Groups (cont.)

- 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.

62

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



63

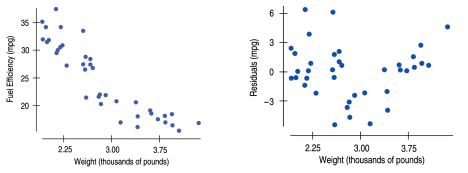
Getting the "Bends"

- 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.

64

Getting the "Bends" (cont.)

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



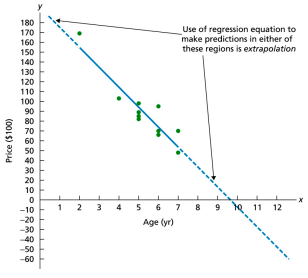
65

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.



66

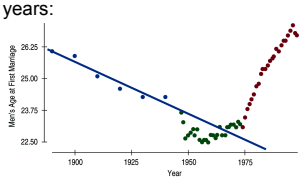
Extrapolation (cont.)

- 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.

67

Extrapolation (cont.)

- A regression of mean age at first marriage for men vs. year fit to the years from 1890 - 1998 does not hold for later years:



- After 1950, linearity did not hold.

68

Predicting the Future

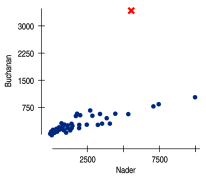
- Extrapolation is always dangerous. But, when the *x*-variable in the model is *Time*, extrapolation becomes an attempt to peer into the future.
- Knowing that extrapolation is dangerous doesn't stop people. The temptation to see into the future is hard to resist.
- Here's some more realistic advice: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

69

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.

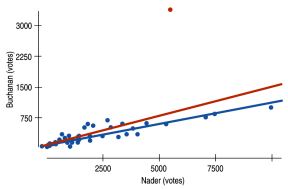
- Any point that stands away from the others can be called an outlier and deserves your special attention.
- The scatterplot shows an outlier.



70

Outliers, Leverage, and Influence (cont.)

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



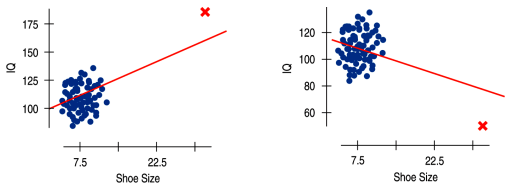
71

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.
- 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.

72

Outliers, Leverage, and Influence (cont.)



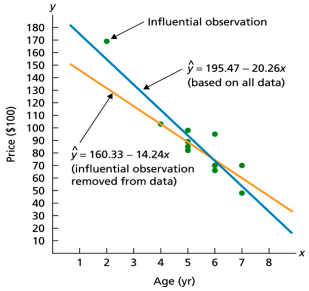
73

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.
- **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.

74

Influential Observation



75

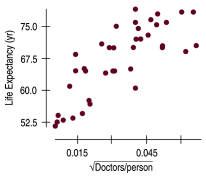
Lurking Variables and Causation

- 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.

76

Lurking Variables and Causation (cont.)

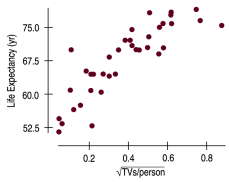
The following scatterplot shows that the average *life expectancy* for a country is related to the number of *doctors* per person in that country:



77

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:



78

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…
  - 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.

79

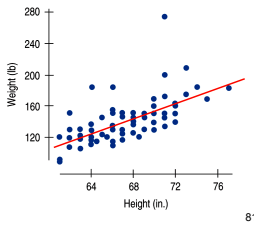
Summary Values

- 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.

80

Working With Summary Values

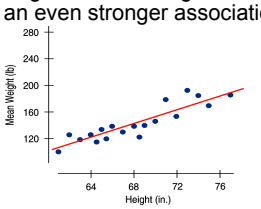
- There is a strong, positive, linear association between *weight* (in pounds) and *height* (in inches) for men:



81

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:



82

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.

83

What have we learned?

We have learned to:

1. Define and apply the concepts related to linear equations with one independent variable.
2. Explain the least-squares criterion.
3. Obtain and graph the regression equation for a set of data points, interpret the slope of the regression line, and use the regression equation to make predictions.
4. Define and use the terminology predictor variable and response variable.
5. Understand the concept of extrapolation.
6. Identify outliers and influential observations.
7. Know when obtaining a regression line for a set of data points is appropriate.

84

What have we learned? (Cont.)

8. Determine and interpret the coefficient of determination.
9. Determine and interpret the linear correlation coefficient, *r*.
10. Explain and apply the relationship between the linear correlation coefficient and the coefficient of determination.

85

Credit

Some of the slides have been adapted/modified in part/whole from the slides of the following textbooks.

- Weiss, Neil A., Introductory Statistics, 8th Edition
- Bock, David E., Stats: Data and Models, 3rd Edition

86