Stat 13, Intro. to Statistical Methods for the Life and Health Sciences.

1. Midterms.
2. Polls.
3. Two quantitative variables, correlation.
4. Linear regression.

No class Thu Nov 24, Thanksgiving.
Read ch10.
Hw4 is 10.1.8, 10.3.14, 10.3.21, 10.4.11 and is due Tue Nov 29.

The final Fri Dec 9, 8am-11, right here, will be on ch1-10.
Bring a PENCIL and CALCULATOR and any books or notes you want. No computers.
http://www.stat.ucla.edu/~frederic/13/F16 .

1. Midterms.
The scores are listed in midtermscores.txt . They are out of 20.

The mean was 16.628 = 83.14%. Median = 85%. SD = 15%.

I do reward improvement on the final.

2. Polls.
a. Why were they so far off?

Before we get into this, when it comes to the election, please make sure to respect freedom of speech and also to respect everyone's right to their beliefs even if they are far from your own.

Our dept's official statement:

In the aftermath of the presidential election, in which passions ran high surrounding issues of tolerance for diversity, it is important to remember that harassment and discrimination based on such things as:

- race, ethnicity, ancestry, color
- sex, gender, gender identity, gender expression, sexual orientation
- national origin, citizenship status
- religion

are not acceptable at UCLA, and may have serious consequences. Information for how to obtain redress or counseling if you are subjected to such harassment or discrimination can be found at:

https://equity.ucla.edu/report-an-incident/

2. Polls.
a. Why were they so far off?

www.realclearpolitics.com/epolls/2016/president/wi/wisconsin_trump_vs_cli

Polling Data								
Poll	Date	Sample	MoE	Clinton (D)	Trump (R)	Johnson (L)	Stein (G)	Spread
Final Results	--	--	--	46.9	47.9	3.6	1.1	Trump +1.0
RCP Average	10/26 - 11/2	--	--	46.8	40.3	5.8	2.0	Clinton +6.5
Remington Research (R)*	11/1 - 11/2	2720 LV	1.9	49	41	3	--	Clinton +8
Loras	10/31 - 11/1	500 LV	4.4	44	38	7	2	Clinton +6
Remington Research (R)*	10/30 - 10/30	1172 LV	2.9	46	42	4	--	Clinton +4
Marquette	10/26 - 10/31	1225 LV	3.5	46	40	4	3	Clinton +6
Emerson	10/26 - 10/27	400 LV	4.9	48	42	9	1	Clinton +6
Remington Research (R)*	10/20 - 10/22	1795 LV	2.3	46	41	5	--	Clinton +5
Monmouth	10/15 - 10/18	403 LV	4.9	47	40	6	1	Clinton +7
WPR/St. Norbert	10/13 - 10/16	644 LV	3.8	47	39	1	3	Clinton +8
Marquette	10/6 - 10/9	878 LV	3.9	44	37	9	3	Clinton +7
CBS News/YouGov	10/5 - 10/7	993 LV	4.3	43	39	4	1	Clinton +4
Loras	10/4 - 10/5	500 LV	4.4	43	35	8	2	Clinton +8
Gravis	10/4 - 10/4	1102 RV	3.0	48	40	4	1	Clinton +8
Emerson	9/19 - 9/20	700 LV	3.6	45	38	11	2	Clinton +7
Marquette	9/15 - 9/18	677 LV	4.8	41	38	11	2	Clinton +3
Monmouth	8/27 - 8/30	404 LV	4.9	43	38	7	3	Clinton +5
Marquette	8/25 - 8/28	650 LV	5.0	41	38	10	4	Clinton +3
Marquette	8/4 - 8/7	683 LV	5.0	47	34	9	3	Clinton +13
Marquette	7/7 - 7/10	665 LV	4.5	43	37	8	2	Clinton +6
CBS News/YouGov*	6/21 - 6/24	993 LV	4.3	41	36	3	2	Clinton +5

2. Polls.

In total this makes 17,104 likely voters in those Wisconsin polls put together.

They averaged **40.3%** for Trump, and Clinton 46.8%. The difference is 6.5%.

Combined, the margin of error for a 95% confidence interval around Trump's percentage would be 0.735%.

The standard error is 0.375% on the estimate of Trump's percentage of 40.3%,

and he got **47.9%**. So they were off by 7.6% which is more than 20

standard errors. The probability is 1 in 10^90 that the polls would be off by that much or more just by chance, if the answers to the polls were just a random sample of how people were actually going to vote.

Technically, there are undecided voters in the polls also. Just taking the difference in percentages between Trump and Hillary Clinton rather than the percentage for Trump into account, the results were off by about 10 SEs, not 20, and this makes the probability of something this extreme or more extreme still astronomical, about $1.5 * 10^{-23}$.

The chance of a monkey randomly typing 15 letters completely at random and happening to choose "hillary r clinton" in order, would be $6 * 10^{-22}$, so it's about 40 times more likely.

What do we conclude?

2. Polls.

What do we conclude?

Either

* lots of people changed their minds,
* the polls were biased,
* the official results were incorrect,

or

* the polls weren't independent of each other.

Those are really the only tenable explanations.

2. Polls.

b. If Clinton had outperformed the polls, what would have been an explanation?
* fundraising advantage.
* more experienced campaign staff.
* more organized ground game.
* Latino populations had surged.
* A higher percentage of Latinos voted.
* Early voting had enabled more poor people and minorities to vote. Thus people who might be considered unlikely voters due to voting trends in previous elections might actually be voting, and the majority of these would be expected to be Democrats.
* Mostly good news for her right before the election.
The FBI said they went through the emails and cleared her of charges.
Lots of big stars were performing and getting people out to the polls and rallies. Obama, Michelle, Bill Clinton, and many others were campaigning hard for her in the final days.
* Meanwhile, many top Republicans were not even supporting Trump. His ground game and field offices were disorganized or nonexistent. Even some right wing radio hosts were criticizing Trump. He had fired his campaign manager midway through the campaign and had an inexperienced hodge podge of supporters and staff.

2. Polls.

c. Other possible pro-Clinton explanations.

* Hillary prepared very early for her run for office. Trump was a latecomer.
* The Democrats mostly coalesced around Hillary, allowing her to pile up numerous endorsements very early. Only a couple of people even ran against her, and none were really promising candidates. Even Sanders was not seen as a very viable candidate when the race began.
* On the Republican side, Trump had to fend off 16 other candidates while Hillary was raising money and holding onto it.
* After Trump got the nomination, barely, he had little convention bounce, and Hillary had a huge one and got a big lead in the polls.
* Obama's popularity has been high.
* The economy, while not too strong, is much much stronger than when Obama began in office, and he was able to campaign strongly for Hillary.
* She also had an incredibly charismatic and great speaker in her husband, and Michelle made great speeches as well.
* The father of a muslim fallen soldier spoke eloquently against Trump.
* Melania got caught blatantly plagiarizing Michelle's speech from 2008.
* Trump got sued for fraud for Trump University, and criticized the judge as unfit because he was of Mexican descent.
* Trump made fun of a reporter in a wheelchair.
* Clinton won all 3 debates according to most polls and surveys.

2. Polls.

c. Other possible pro-Clinton explanations.
* Trump continued to refuse to release his tax returns despite prodding.
* The tape came out where Trump bragged about grabbing women. Clinton surged way ahead in the polls.
* Many important Republicans stopped supporting him.
* Down ballot Republicans tried to distance themselves from him.
* Clinton was raising more funds than Trump, and Trump was not spending much of his own money on his campaign.

d. Bayesian statistics, Nate Silver, and 538.com.

In his most recent article on 538.com, Nate Silver states very clearly that he expected Hillary to win, and his models favored her, but he also emphasizes that the polls are typically off by the amounts they were off this time and points out that his models gave Trump about a 30% chance of winning the day before the election.

He also admits that the polls typically miss by amounts *greater than what we would expect due to sampling error alone*.

d. Bayesian statistics, Nate Silver, and 538.com.

As Nate Silver notes, the polls indicated a large proportion of undecided voters, which meant there was more uncertainty and therefore a larger chance for a major polling error.

Nate Silver and 538.com use a Bayesian model to forecast the election.
How does Bayesian modeling work?
The model starts with *prior distributions* which are supposed to reflect the researcher's beliefs about the probability of something before collecting data.
For instance, you might have a prior distribution that the percentage μ of votes the Democratic candidate would get is spread uniformly between 40% and 60%.
Then you collect data, from polls, economic data, etc., and gradually update your distribution, forming a *posterior distribution* for μ.
Nate Silver's main idea was to weight the different polls by how accurate they were historically.

Bayesian modelers are often correctly criticized for not adequately validating their models.
Nate Silver might be a good modeler, but it is hard to tell.

d. Bayesian statistics, Nate Silver, and 538.com.

"Apart from calibration, are there other good methods to evaluate a probabilistic forecast? Not really, although sometimes it can be worthwhile to look for signs of whether an upset winner benefited from good luck or quirky, one-off circumstances." Nate Silver, May 18, 2016.
http://fivethirtyeight.com/features/how-i-acted-like-a-pundit-and-screwed-up-on-donald-trump

Calibration. If we take all candidates where the model outputs a probability of 35%, do the candidates win 35% of the time?

I could just say every Democrat and Republican has a 50% chance every election. This model might be well calibrated, but it is not very informative.
Validation. Is the model accurate? Does it outperform competing models?
e.g. Log-likelihood: $\sum_{i\ won} \log(p_i) + \sum_{i\ lost} \log(1-p_i)$.

There are potential limitations of Bayesian statistics.
* Why might one want to know if Clinton's probability of winning is 67% or 72%? Perhaps if one is only very superficially interested?
* What if my prior is different from Nate Silver's?
* If you are deeply interested, why not look at the poll results themselves, directly?
* Bayesian models might be more difficult to validate, in some cases.

e. Are there other predictors of success in Presidential elections?

Here are the presidential elections that have occurred in my lifetime.

1972. Nixon vs. McGovern.

1976. Carter vs. Ford.

1980. Reagan vs. Carter.

1984. Reagan vs. Mondale.

1988. Bush vs. Dukakis.

1992. Clinton vs. Bush.

1996. Clinton vs. Dole.

2000. Bush vs. Gore.

2004. Bush vs. Kerry.

2008. Obama vs. McCain.

2012. Obama vs. Romney.

2016. Trump vs. Clinton.

Many would say the candidate with more humor, style, and charisma won 11/12 of these elections. The two-sided p-value = 0.63%.

Experience? Hard to say, but perhaps the more experienced candidate won 3/12. The two-sided p-value = 14.6%.

3. Two Quantitative Variables

Chapter 10

Two Quantitative Variables: Scatterplots and Correlation

Section 10.1

Scatterplots and Correlation

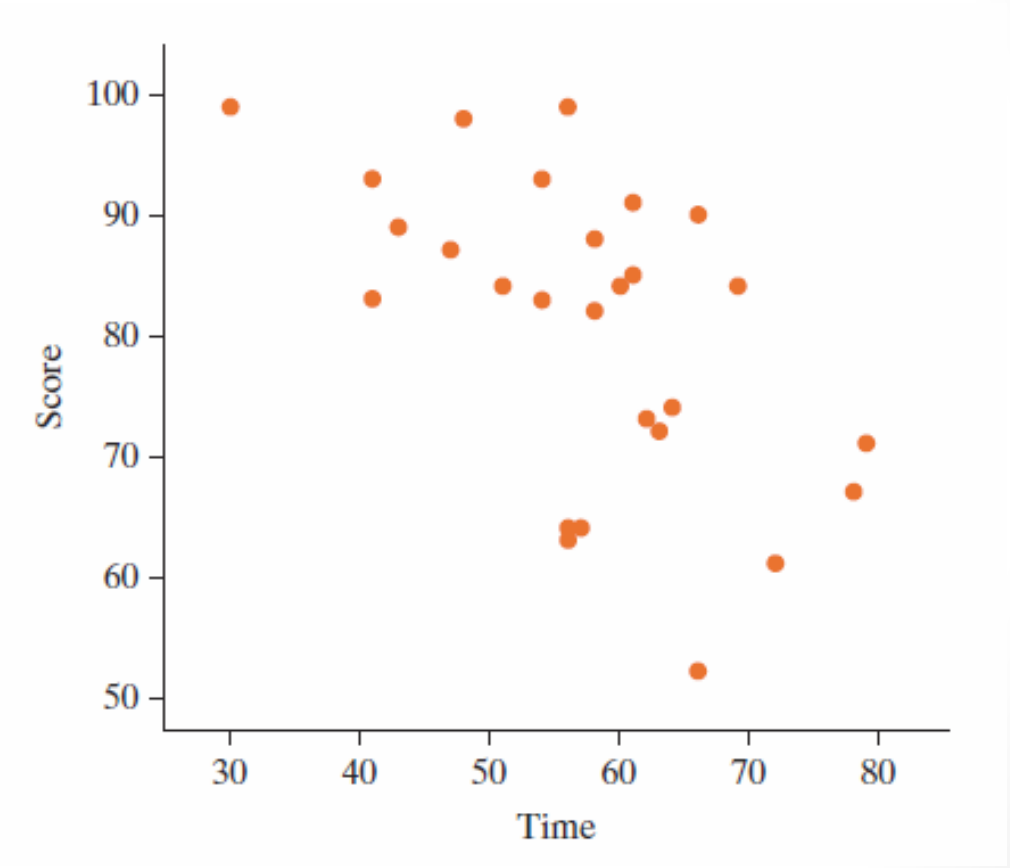
Suppose we collected data on the relationship between the time it takes a student to take a test and the resulting score.

Time	30	41	41	43	47	48	51	54	54	56	56	56	57	58
Score	100	84	94	90	88	99	85	84	94	100	65	64	65	89
Time	58	60	61	61	62	63	64	66	66	69	72	78	79	
Score	83	85	86	92	74	73	75	53	91	85	62	68	72	

Scatterplot

Put explanatory variable on the horizontal axis.

Put response variable on the vertical axis.



Describing Scatterplots

- When we describe data in a scatterplot, we describe the
  - Direction (positive or negative)
  - Form (linear or not)
  - Strength (strong-moderate-weak, we will let correlation help us decide)
  - Unusual Observations
- How would you describe the time and test scatterplot?

Correlation

- **Correlation** measures the strength and direction of a linear association between two quantitative variables.
- Correlation is a number between -1 and 1.
- With positive correlation one variable increases, on average, as the other increases.
- With negative correlation one variable decreases, on average, as the other increases.
- The closer it is to either -1 or 1 the closer the points fit to a line.
- The correlation for the test data is -0.56.

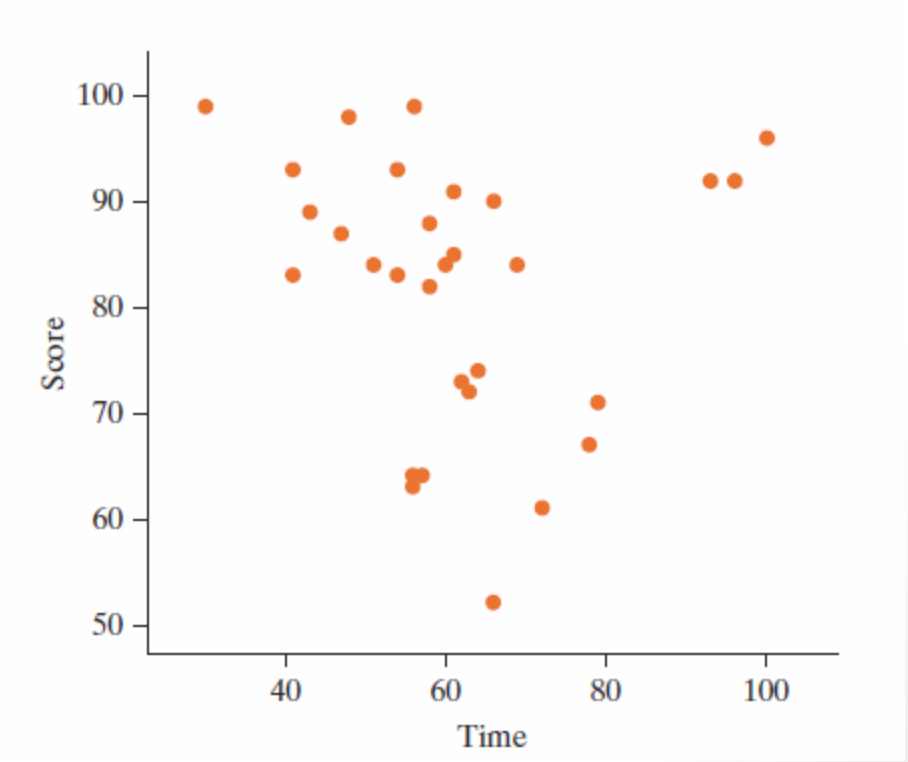
Correlation Guidelines

Correlation Value	Strength of Association	What this means
0.7 to 1.0	Strong	The points will appear to be nearly a straight line
0.3 to 0.7	Moderate	When looking at the graph the increasing/decreasing pattern will be clear, but there is considerable scatter.
0.1 to 0.3	Weak	With some effort you will be able to see a slightly increasing/decreasing pattern
0 to 0.1	None	No discernible increasing/decreasing pattern
Same Strength Results with Negative Correlations		

Back to the test data

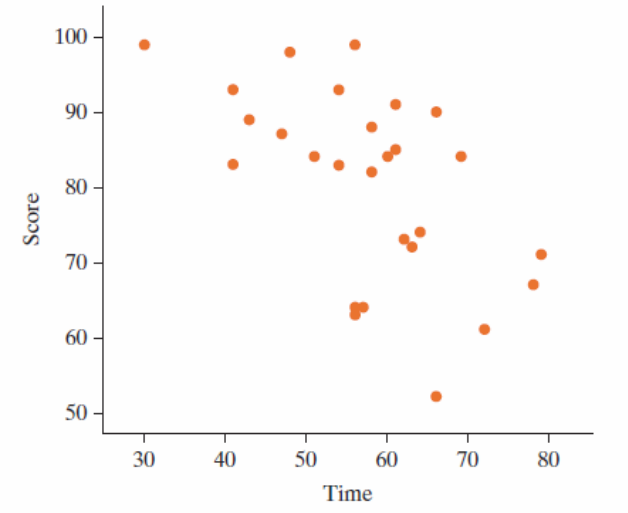
Actually the last three people to finish the test had scores of 93, 93, and 97.

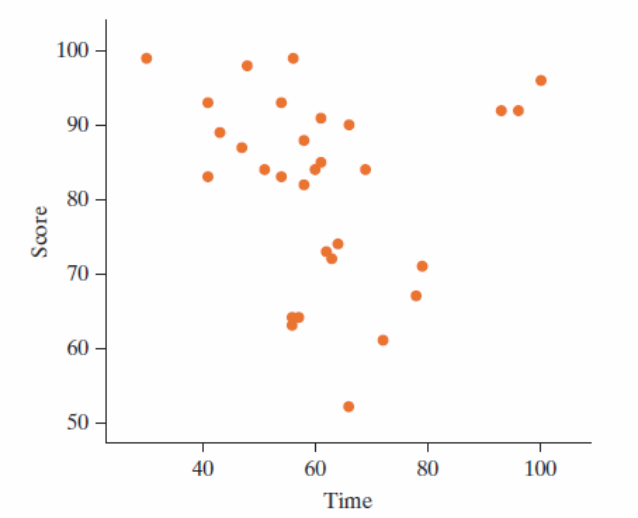
What does this do to the correlation?



Influential Observations

- The correlation changed from -0.56 (a fairly moderate negative correlation) to -0.12 (a weak negative correlation).
- Points that are far to the left or right and not in the overall direction of the scatterplot can greatly change the correlation. (influential observations)





Correlation

- **Correlation** measures the strength and direction of a linear association between two quantitative variables.
  - $-1 \leq r \leq 1$
  - Correlation makes no distinction between explanatory and response variables.
  - Correlation has no units.
  - Correlation is not resistant to outliers. It is sensitive.

Learning Objectives for Section 10.1

- Summarize the characteristics of a scatterplot by describing its direction, form, strength and whether there are any unusual observations.
- Recognize that the correlation coefficient is appropriate only for summarizing the strength and direction of a scatterplot that has linear form.
- Recognize that a scatterplot is the appropriate graph for displaying the relationship between two quantitative variables and create a scatterplot from raw data.
- Recognize that a correlation coefficient of 0 means there is no linear association between the two variables and that a correlation coefficient of -1 or 1 means that the scatterplot is exactly a straight line.
- Understand that the correlation coefficient is influenced by extreme observations.

Inference for the Correlation Coefficient: Simulation-Based Approach

Section 10.2

We will look at a small sample example to see if body temperature is associated with heart rate.

Temperature and Heart Rate

Hypotheses

- Null: There is no association between heart rate and body temperature. ($\rho = 0$)
- Alternative: There is a positive linear association between heart rate and body temperature. ($\rho > 0$)

ρ = rho

Inference for Correlation with Simulation (Section 10.2)

1. Compute the observed statistic. (Correlation)
2. Scramble the response variable, compute the simulated statistic, and repeat this process many times.
3. Reject the null hypothesis if the observed statistic is in the tail of the null distribution.

Temperature and Heart Rate

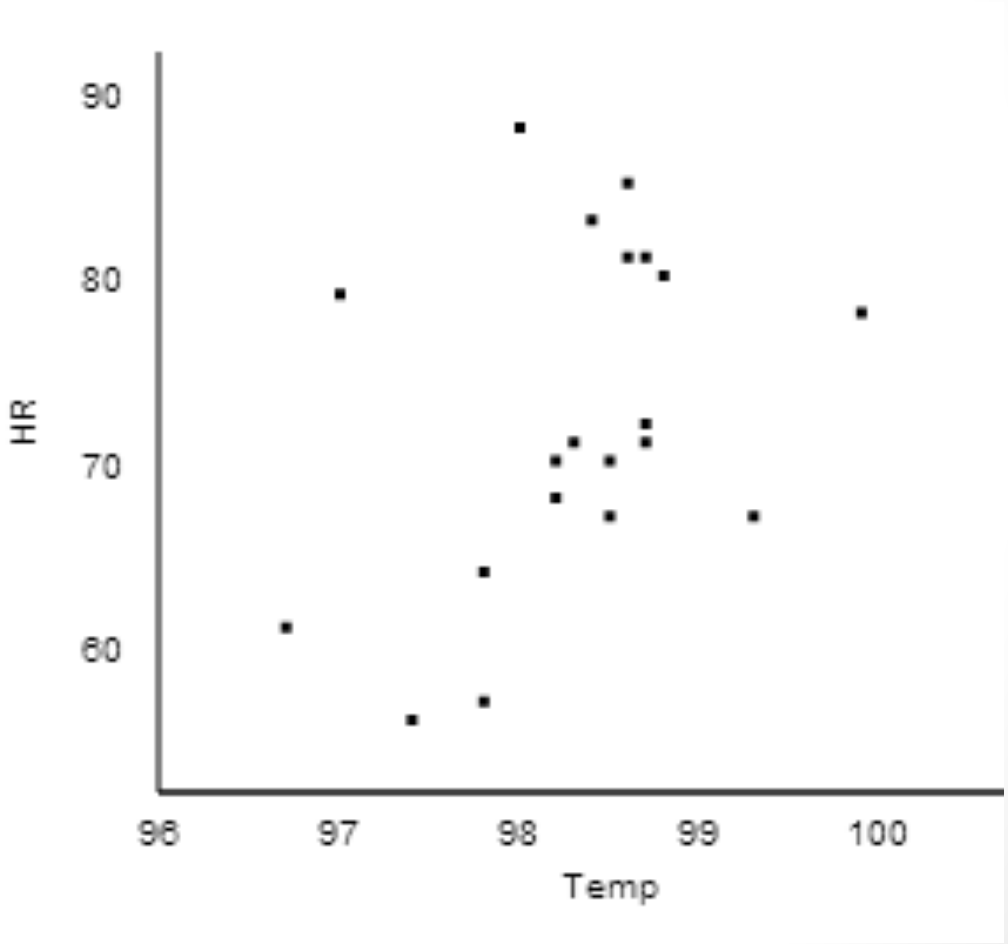
Collect the Data

Tmp	98.3	98.2	98.7	98.5	97.0	98.8	98.5	98.7	99.3	97.8
HR	72	69	72	71	80	81	68	82	68	65
Tmp	98.2	99.9	98.6	98.6	97.8	98.4	98.7	97.4	96.7	98.0
HR	71	79	86	82	58	84	73	57	62	89

Temperature and Heart Rate

Explore the Data

r = 0.378



Temperature and Heart Rate

- If there was no association between heart rate and body temperature, what is the probability we would get a correlation as high as 0.378 just by chance?

- If there is no association, we can break apart the temperatures and their corresponding heart rates. We will do this by shuffling one of the variables.

Shuffling Cards

- Let's remind ourselves what we did with cards to find our simulated statistics.
- With two proportions, we wrote the response on the cards, shuffled the cards and placed them into two piles corresponding to the two categories of the explanatory variable.
- With two means we did the same thing except this time the responses were numbers instead of words.

Dolphin Therapy

Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver
Non-improver	Improver	Improver

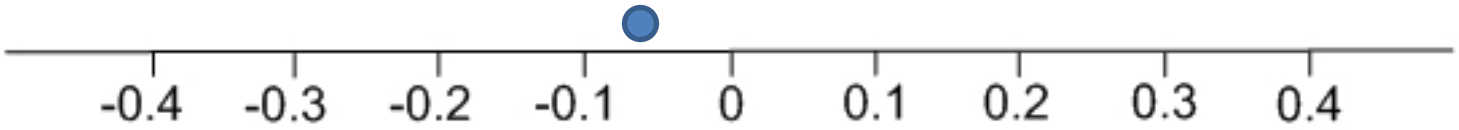
40.0% Improvers

Control

Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Non-improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver
Non-improver	Non-improver	Improver

46.7% Improvers

0.400 – 0.467 = -0.067



Difference in Simulated Proportions

Music

25.2 45.6
14.5 11.6
-7.0 18.6
12.6 12.1
34.5 30.5

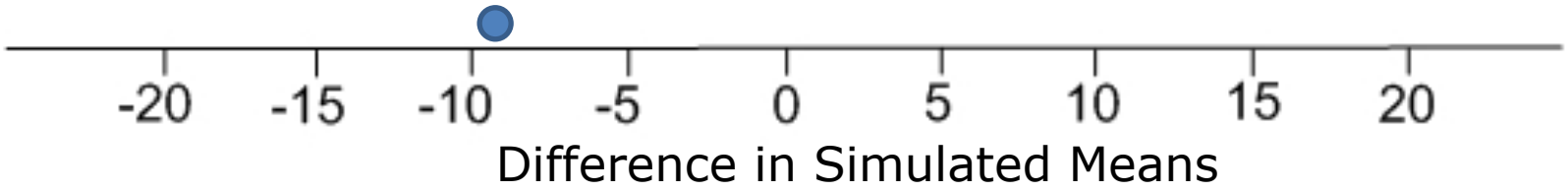
mean = 16.12

No music

-10.7 -10.7 10.0
4.5 9.6
2.2 2.4
21.3 21.8
-14.7 7.2

mean = 6.38

6.38 – 16.12 = -9.74



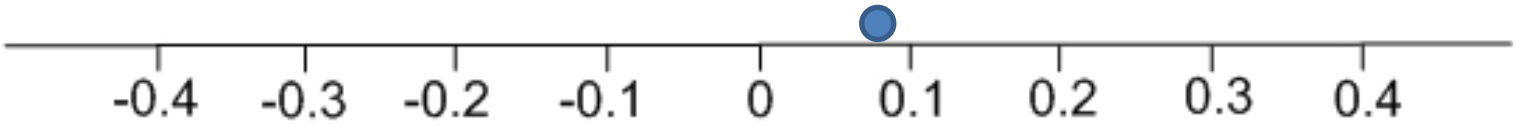
Shuffling Cards

- Now how will this shuffling be different when both the response and the explanatory variable are quantitative?
- We can’t put things in two piles anymore.
- We still shuffle values of the response variable, but this time place them next to two values of the explanatory variable.

Body Temperature and Heart Rate

98.3°	98.2°	97.7°	98.5°	97.0°	98.8°	98.5°	98.7°	99.3°	97.8°
72	69	72	71	80	81	68	82	68	65
98.2°	99.9°	98.6°	98.6°	97.8°	98.4°	98.7°	97.4°	96.7°	98.0°
71	79	86	82	58	84	73	57	62	89

r = 0.078

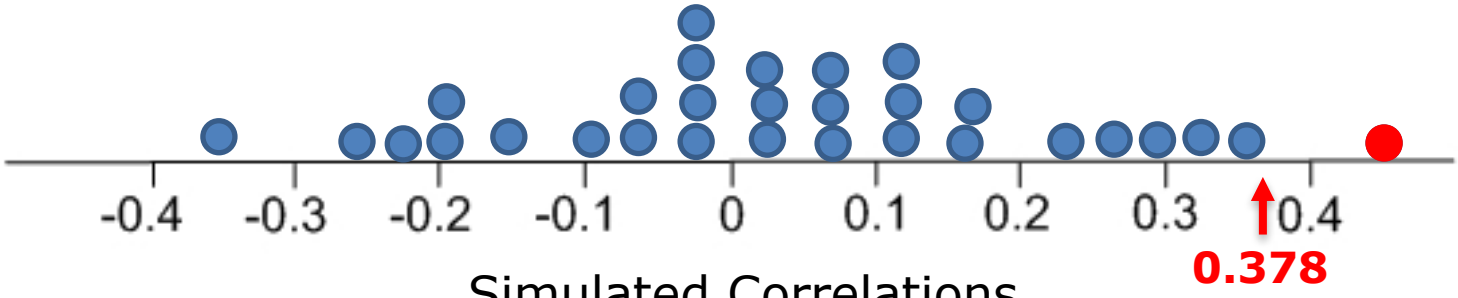


Simulated Correlations

More Simulations

0.097 0.054 0.062 0.259 0.034
-0.253 -0.345 0.339
0.314
0.059 0.200 0.447
0.042 0.167
0.067
0.232 -0.327 0.100 0.329

Only one simulated statistic out of 30 was as large or larger than our observed correlation of 0.378, hence our p-value for this null distribution is 1/30 ≈ 0.03.

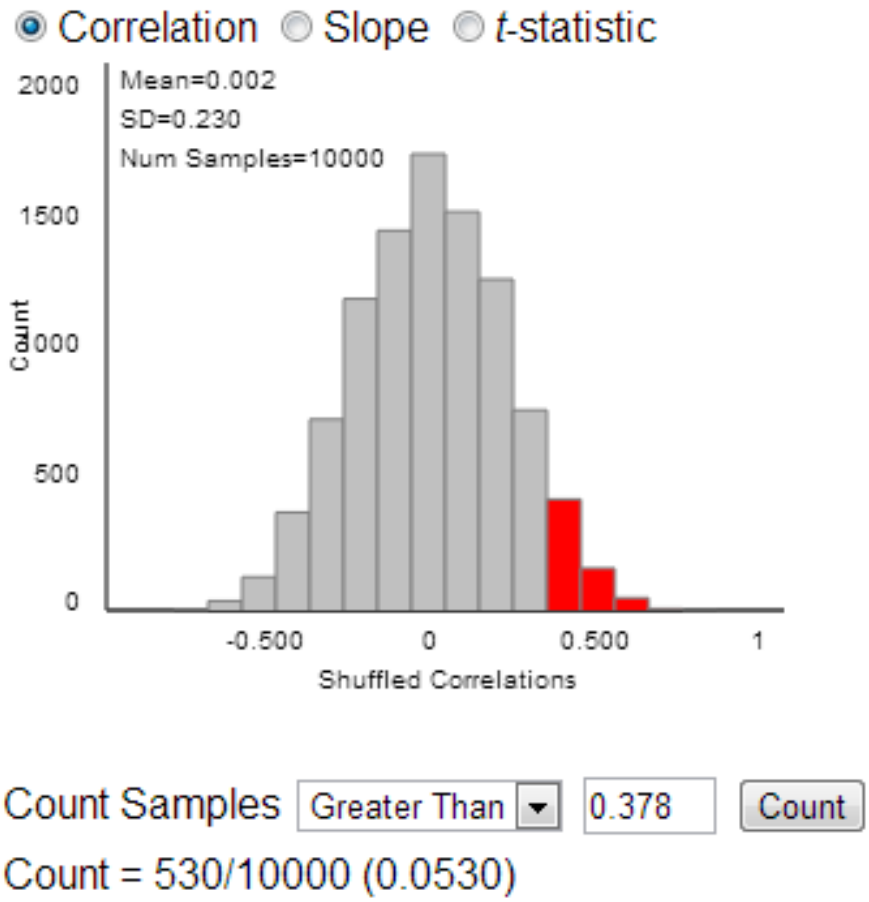


Temperature and Heart Rate

- We can look at the output of 1000 shuffles with a distribution of 1000 simulated correlations.

Temperature and Heart Rate

- Notice our null distribution is centered at 0 and somewhat symmetric.
- We found that 530/10000 times we had a simulated correlation greater than or equal to 0.378.



Temperature and Heart Rate

- With a p-value of 0.053 = 5.3%, we almost but do not quite have statistical significance. This is moderate evidence of a positive linear association between body temperature and heart rate. Perhaps a larger sample would give a smaller p-value.

4. Least Squares Regression

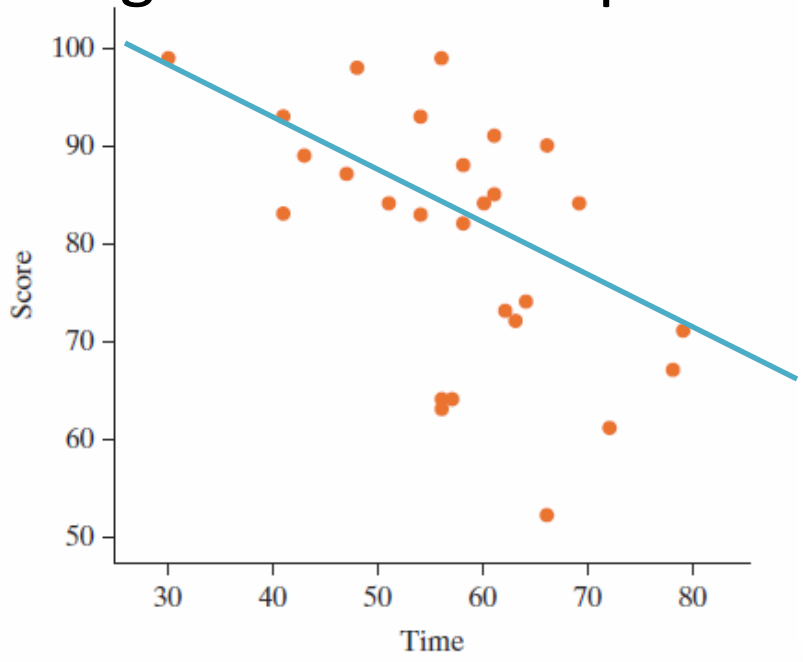
Section 10.3

Introduction

- If we decide an association is linear, it is helpful to develop a mathematical model of that association.
- Helps make predictions about the response variable.
- The ***least-squares regression line*** is the most common way of doing this.

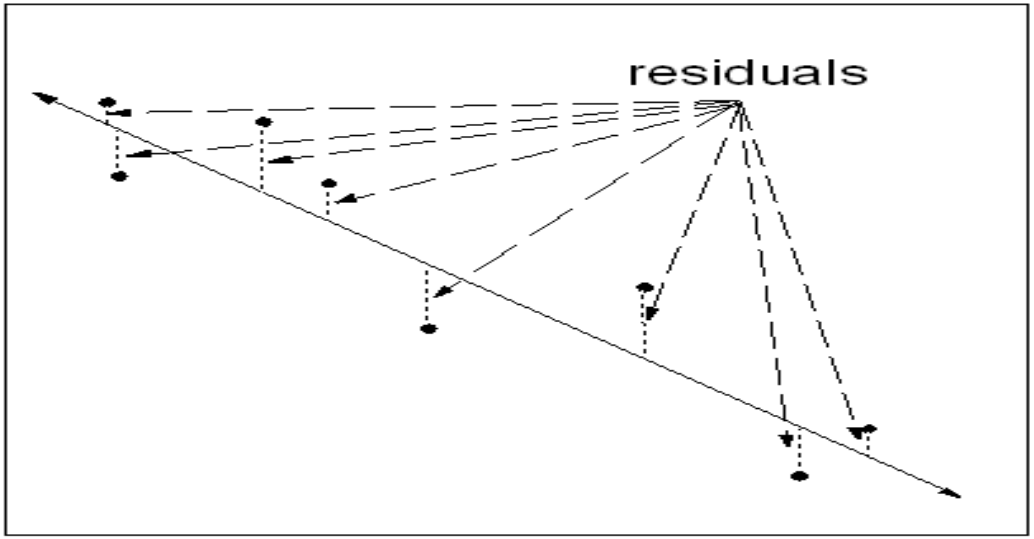
Introduction

- Unless the points are perfectly linearly alligned, there will not be a single line that goes through every point.
- We want a line that gets as close as possible to all the points.



Introduction

- We want a line that minimizes the vertical distances between the line and the points
  - These distances are called **residuals.**
  - The line we will find actually minimizes the sum of the squares of the residuals.
  - This is called a **least-squares regression line**.

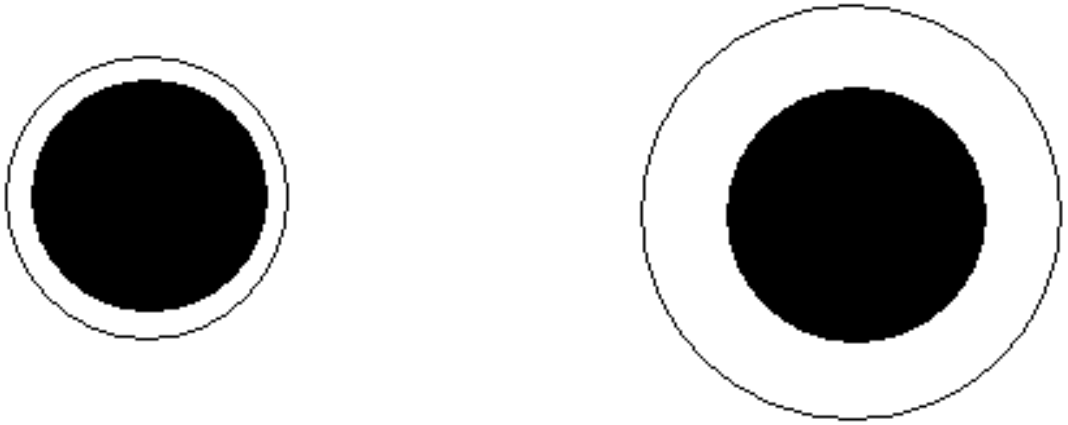


Are Dinner Plates Getting Larger?

Example 10.3

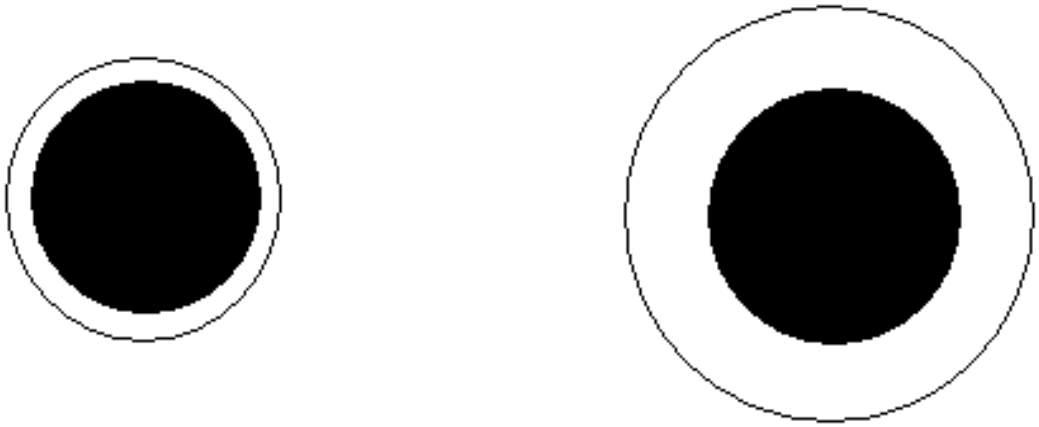
Growing Plates?

- There are many recent articles and TV reports about the obesity problem.
- One reason some have given is that the size of dinner plates are increasing.
- Are these black circles the same size, or is one larger than the other?



Growing Plates?

- They appear to be the same size for many, but the one on the right is about 20% larger than the left.



- This suggests that people will put more food on larger dinner plates without knowing it.
- There is name for this phenomenon: *Delboeuf illusion*

Growing Plates?

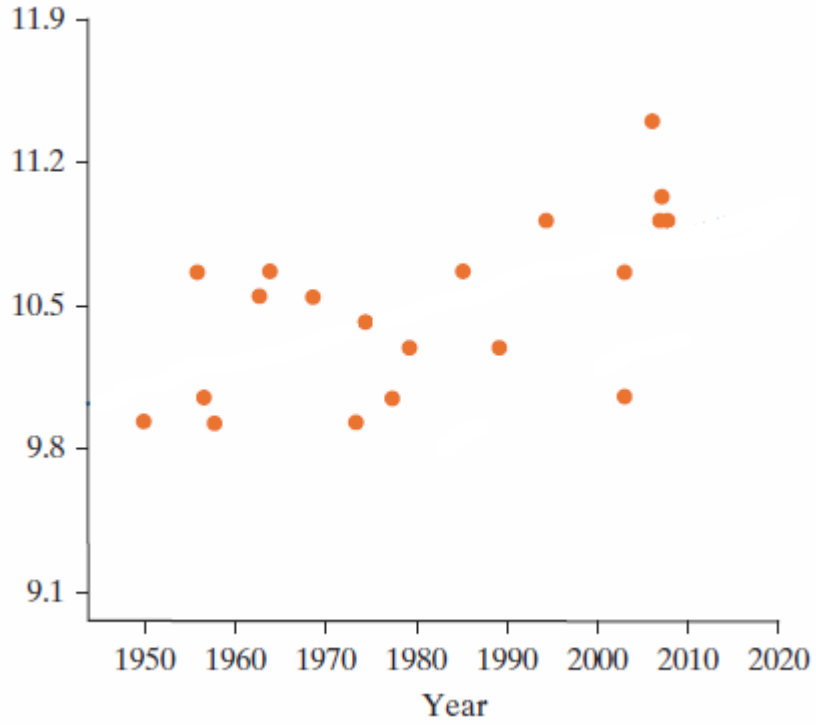
- Researchers gathered data to investigate the claim that dinner plates are growing
- American dinner plates sold on ebay on March 30, 2010 (Van Ittersum and Wansink, 2011)
- Year manufactured and diameter are given.

TABLE 10.1 Data for size (diameter, in inches) and year of manufacture for 20 American-made dinner plates

Year	1950	1956	1957	1958	1963	1964	1969	1974	1975	1978
Size	10	10.75	10.125	10	10.625	10.75	10.625	10	10.5	10.125
Year	1980	1986	1990	1995	2004	2004	2007	2008	2008	2009
Size	10.375	10.75	10.375	11	10.75	10.125	11.5	11	11.125	11

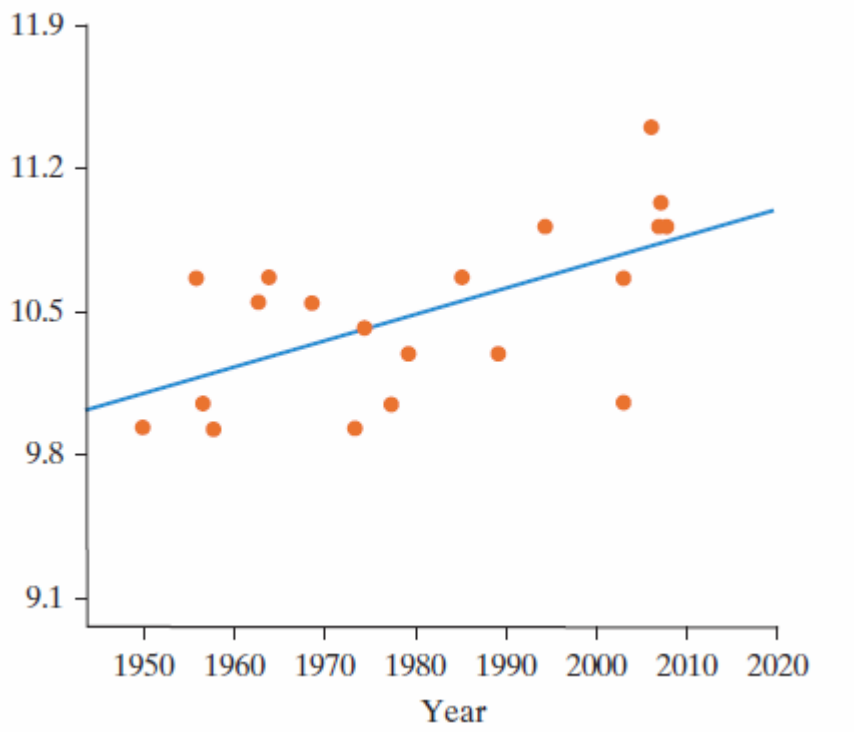
Growing Plates?

- Both year (explanatory variable) and diameter in inches (response variable) are quantitative.
- Each dot represents one plate in this scatterplot.
- Describe the association here.



Growing Plates?

- The association appears to be roughly linear
- The least squares regression line is added
- How can we describe this line?



Regression Line

The regression equation is $\hat{y} = a + bx$:

- *a* is the *y*-intercept
- *b* is the slope
- *x* is a value of the explanatory variable
- $\hat{y}$ is the predicted value for the response variable

- For a specific value of *x*, the corresponding distance $y - \hat{y}$ (or actual – predicted) is a residual

Regression Line

- The least squares line for the dinner plate data is $\hat{y} = -14.8 + 0.0128x$
- Or $\widehat{\text{diameter}} = -14.8 + 0.0128(\text{year})$
- This allows us to predict plate diameter for a particular year.

Slope

$$\hat{y} = -14.8 + 0.0128x$$

- What is the predicted diameter for a plate manufactured in 2000?
  - -14.8 + 0.0128(2000) = 10.8 in.
- What is the predicted diameter for a plate manufactured in 2001?
  - -14.8 + 0.0128(2001) = 10.8128 in.
- How does this compare to our prediction for the year 2000?
  - 0.0128 larger
- Slope *b* = 0.0128 means that diameters are predicted to increase by 0.0128 inches per year on average

Slope

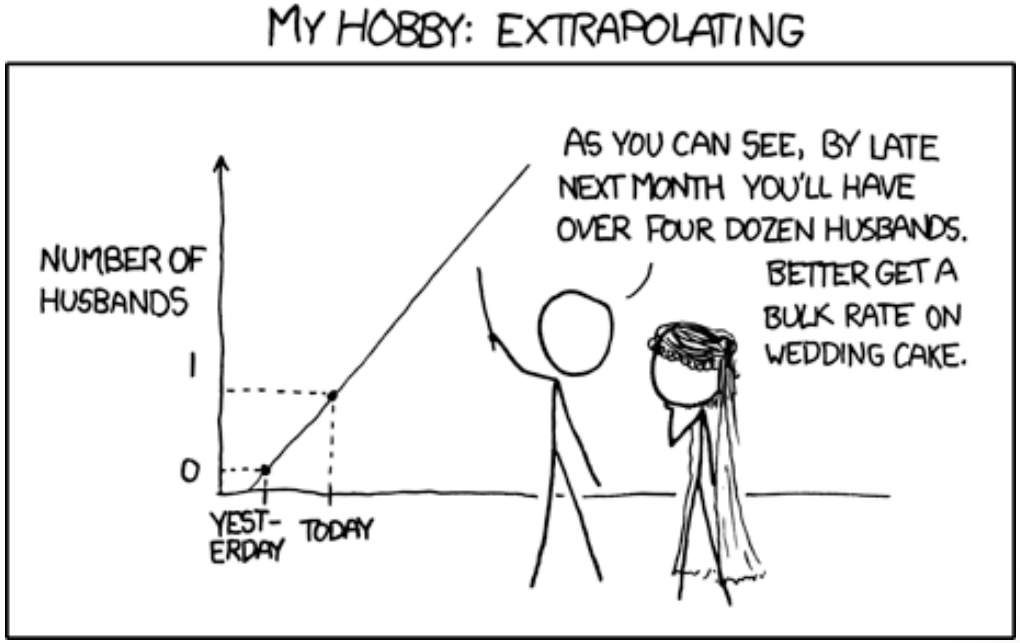
- Slope is the predicted change in the response variable for one-unit change in the explanatory variable.
- Both the slope and the correlation coefficient for this study were positive.
  - The slope is 0.0128
  - The correlation is 0.604
- The slope and correlation coefficient will always have the same sign.

y-intercept

- The *y*-intercept is where the regression line crosses the *y*-axis or the predicted response when the explanatory variable equals 0.
- We had a *y*-intercept of -14.8 in the dinner plate equation. What does this tell us about our dinner plate example?
  - Dinner plates in year 0 were -14.8 inches.
- How can it be negative?
  - The equation works well within the range of values given for the explanatory variable, but fails outside that range.
- Our equation should only be used to predict the size of dinner plates from about 1950 to 2010.

Extrapolation

- Predicting values for the response variable for values of the explanatory variable that are outside of the range of the original data is called ***extrapolation***.



Coefficient of Determination

- While the intercept and slope have meaning in the context of year and diameter, remember that the correlation does not. It is just 0.604.
- However, the square of the correlation (coefficient of determination or r^2) does have meaning.
- $r^2 = 0.604^2 = 0.365$ or 36.5%
- 36.5% of the variation in plate size (the response variable) can be explained by its linear association with the year (the explanatory variable).

Learning Objectives for Section 10.3

- Understand that one way a scatterplot can be summarized is by fitting the best-fit (least squares regression) line.
- Be able to interpret both the slope and intercept of a best-fit line in the context of the two variables on the scatterplot.
- Find the predicted value of the response variable for a given value of the explanatory variable.
- Understand the concept of residual and find and interpret the residual for an observational unit given the raw data and the equation of the best fit (regression) line.
- Understand the relationship between residuals and strength of association and that the best-fit (regression) line this minimizes the sum of the squared residuals.

Learning Objectives for Section 10.3

- Find and interpret the coefficient of determination (r^2) as the squared correlation and as the percent of total variation in the response variable that is accounted for by the linear association with the explanatory variable.
- Understand that extrapolation is when a regression line is used to predict values outside of the range of observed values for the explanatory variable.
- Understand that when slope = 0 means no association, slope < 0 means negative association, slope > 0 means positive association, and that the sign of the slope will be the same as the sign of the correlation coefficient.
- Understand that influential points can substantially change the equation of the best-fit line.