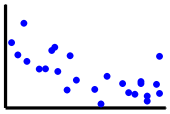


TOPIC D7: RELATIONSHIPS - SUMMARIZING BY A LINE

SPOTLIGHT: MEASURING A CAR

Every fall, Consumers Reports publishes a magazine *New Car Preview* that reviews the new cars that are available for a particular model year. The intent of this publication is to present information about the new cars so you can make an informed decision when you are in the market for buying a car. Specifically, Consumer Reports presents a profile of a particular car model that contains much information and test data about the car. But actually how does this consumer organization measure a car?

One section of measurements about a car is labeled "Specifications." This section contains physical measurements, such as its length (inches), width (inches) and curb weight (pounds). This section also contains important fuel measurements such as its estimated mileage (in miles per gallon) in city and highway driving, and the number of gallons in its fuel tank. Another section of the profile is labeled "From the Test Track." This section presents measurements about the car that were made when Consumer Reports took the car on a test drive. On this test drive, measurements were made on the seating dimensions of the car such as the inches of rear leg room and the inches of front head room. Also in the test drive, Consumer Reports made several measurements of the acceleration and the distance required to brake on a dry surface. Measurements were made on the size of the cargo area in the trunk (in cubic feet) and the maximum load (in pounds) that the car can carry including passengers and cargo. To check the fuel economy, the mileage of the car was measured both in city and highway driving, and the annual fuel cost was estimated.

Above we have focused on quantitative car measurements. But the profile also presents many categorical measurements about the car. These measurements include ratings on the car's reliability with respect to the engine, cooling, fuel, transmission, etc. Also Consumer Reports shows the results on the car's crash tests (from good to bad) and

rates the convenience and comfort of the car such as the ease in using the controls, the front-seat comfort, and the noise level.

Are all of these car measurements helpful when you make your decision on what car to purchase? Probably not. But there will likely be certain measurements that will be important to you when you compare car models. In the case of our family, we wanted to make sure that the car would fit into our garage, so the length of the car was an important measurement. Also the fuel economy measurements were important due to the current high price of gasoline.

Later in this topic, we will explore the relationships between several different car measurements as reported in this magazine.

PREVIEW

In the last topic, we were introduced to the scatterplot that graphically shows the association between two quantitative variables. In this topic we describe ways of summarizing this association. We first introduce a correlation coefficient that measures the direction and strength of the association that we see in the scatterplot. When there is an association, it is helpful to fit a line to the points and we'll describe two methods for fitting this line.

In this topic your learning objectives are to:

- Understand how one can use a line to describe a relationship and predict values of one variable given values of a second variable.
- Understand two methods of fitting a line, and when one method may be preferable to the second method when there are outliers in the data.
- Understand the computation and interpretation of residuals.
- Understand the limitations of a best fitting line in understanding a relationship.
- Understand three ways of studying the relationship between two quantitative variables.
- Understand that the best way of studying the relationship can depend on the particular dataset or way we communicate the relationship.

NCTM Standards

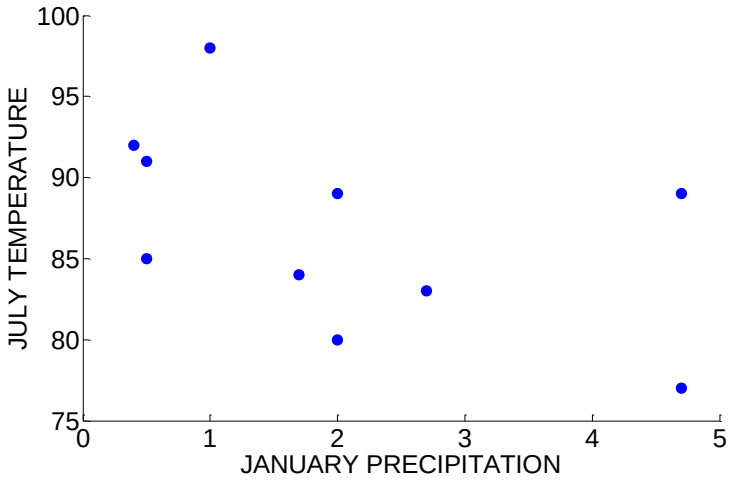
✓ In Grades 6-8, all students should make conjectures about possible relationships between two characteristics of a sample on the basis of scatterplots of the data and approximate lines of fit.

✓In Grades 9-12, all students should, for bivariate measurement data, be able to determine regression coefficients and regression equations using technological tools.

✓In Grades 9-12, all students should identify trends in bivariate data and find functions that model the data.

RELATIONSHIPS - SUMMARIZING BY A LEAST-SQUARES LINE

Again we look at the relationship average amount of precipitation (inches of rain or melted snow) in January and the average high temperature July for ten cities.



In the first part of this topic, we talked about measuring the strength of the association between two measurement variables by means of a correlation coefficient. Suppose we are primarily interested in one variable, called the *response variable* and we wish to use a second variable, called the *explanatory variable*, to predict values of the

response variable. When two variables have a strong association, it is not enough to just state a correlation value – we like to know *how* the response variable changes as we change the explanatory variable. A simple way to describe this relationship is to fit a "best line" to the scatterplot, and use this line to make our predictions.

Predicting a city's average July temperature

Let's suppose that you are given the average July temperatures for the ten cities and you are interested in predicting the average temperature in July for another city, say Metropolis. Can you predict Metropolis' average July temperature? If you were not given any information about this city, then it would be reasonable to assume that Metropolis is similar to the other ten cities with respect to temperature, and so you can predict Metropolis' July temperature by using the mean July temperature for the ten cities. This mean temperature turns out to be 86.8 degrees.

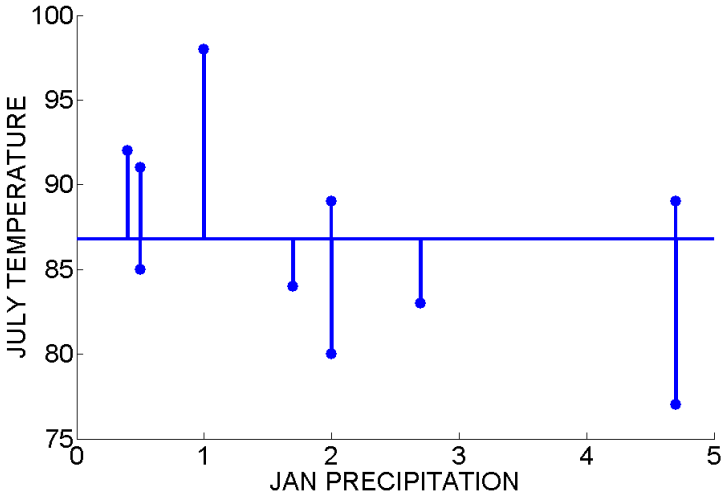
Is this a good prediction? Let's use this prediction, 86.8, to estimate the July temperature for the 10 cities. In the below scatterplot, a horizontal line is drawn at the value 86.8 degrees. This line

$$predicted\ \ July\ temperature = 86.2$$

represents our predicted July temperature. Of course, the cities' actual July temperatures are either smaller or larger than 86.8 degrees. We have drawn vertical lines from the actual July temperatures (the dots) to the predicted values (the horizontal line). We define a residual to be the difference between the actual and predicted temperature:

$$RESIDUAL = Observed\ temperature - Predicted\ temperature\ .$$

The lengths of these lines represent our errors in using 86.8 to predict the July temperatures for the 10 cities. One way of measuring the accuracy of our prediction is by the sum of the squared residuals.



We show how to compute this measure in the below table. We show the July temperature for each city, the predicted value, the residual, and the square of the residual.

City	January temp	July temp	Residual	Squared Residual
Aberdeen	20	85	-1.8	3.24
Akron	33	83	-3.8	14.44
Albuquerque	47	92	5.2	27.04
Amarillo	49	91	4.2	17.64
Aspen	34	80	-6.8	46.24
Atlanta	52	89	2.2	4.84
Bakersfield	57	98	11.2	125.44
Bar Harbor	33	77	-9.8	96.04
Chicago	29	84	-2.8	7.84
Miami	76	89	2.2	4.84
			SUM	347.6

We see from the table that the sum of the squared residuals is 347.6. This sum represents the total size of the error in using the single value, 86.8, to predict the July temperature for the ten cities.

Using January precipitation to predict July temperature

How can we get a better prediction of Metropolis' July temperature? We know from the previous discussion that a city's average January precipitation is negatively associated with a city's July temperature. So maybe if we knew Metropolis' January precipitation, then we could use this information to obtain a better prediction of the city's July temperature.

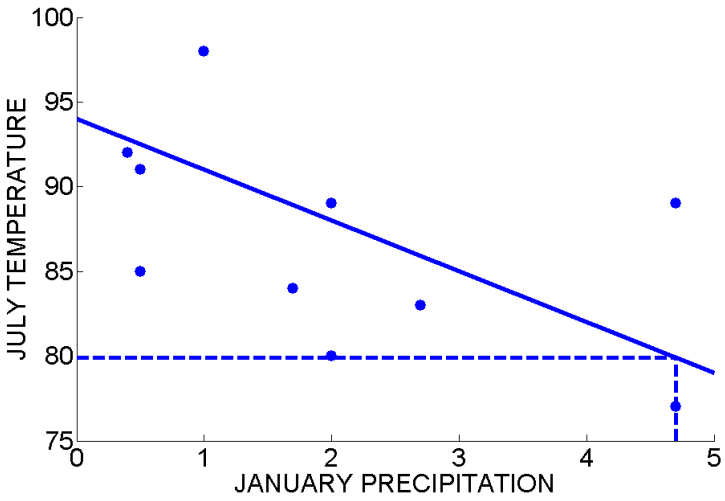
We can find a "good" predictor by fitting a line to our scatterplot. Below we have redrawn the scatterplot of the (Jan precipitation, July temperature) data. Suppose we draw the line

$$predicted\ \ July\ temperature = -3 \times (January\ \ precipitation) + 94$$

on top of the graph. (I found this line by trial and error. I wanted to find a line with a simple formula that went through the middle of the cluster of points in the scatterplot.) This formula gives us one possible prediction for a city's July temperature if we know its January precipitation. For example, Atlanta's average January precipitation is 4.7 inches. Using this line formula, we would predict Atlanta's average July temperature to be

$$-3 \times 4.7 + 94 = 79.9.$$

So we would predict Atlanta's July temperature to be 79.9 degrees. This prediction is shown graphically on the figure.



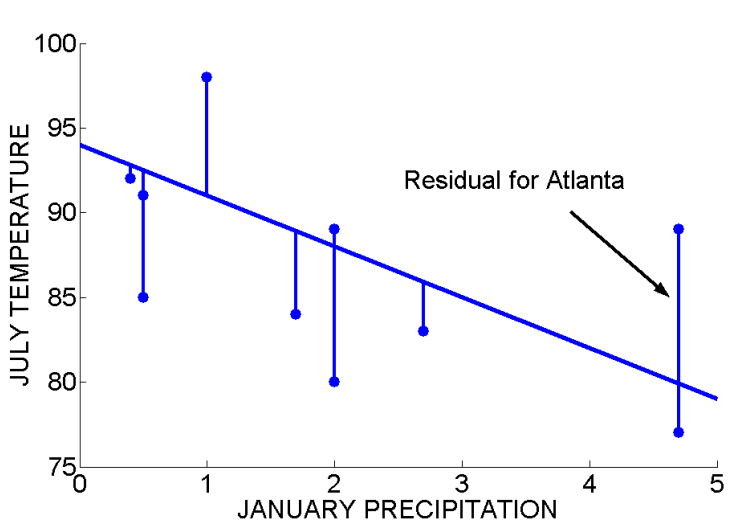
But how good is our prediction? Suppose we use this line formula

$$predicted\ July\ temperature = -3 \times (January\ precipitation) + 94$$

to predict the July temperature for all ten cities. For each city, we can compute a corresponding residual. For example, for Atlanta, our prediction was 79.9 degrees and Atlanta's actual July temperature is 89 degrees. The corresponding residual is

RESIDUAL = Actual July temperature - Predicted July temperature = 89 – 79.9 = 9.1.

On the scatterplot, the residuals are represented by the vertical lines from the observed points to the line. We have pointed out the residual for Atlanta in the figure. For this city, the large positive residual indicates that Atlanta's actual July temperature is far above the line that represents its predicted temperature.



In the table below, we compute the predicted July temperatures for all 10 cities. As in the earlier table, we compute the residuals and the squared residuals.

City	Jan precip	July temp	predicted	Residual	Squared residual
Aberdeen	0.5	85	92.5	-7.5	56.25
Akron	2.7	83	85.9	-2.9	8.41
Albuquerque	0.4	92	92.8	-0.8	0.64
Amarillo	0.5	91	92.5	-1.5	2.25
Aspen	2	80	88	-8	64
Atlanta	4.7	89	79.9	9.1	82.81
Bakersfield	1	98	91	7	49
Bar Harbor	4.7	77	79.9	-2.9	8.41
Chicago	1.7	84	88.9	-4.9	24.01
Miami	2	89	88	1	1
				SUM	296.78

As before, we can measure the accuracy of this line prediction by the sum of squared residuals. Here this sum is 296.78. Recall that the sum of squared residuals for the first line

$$predicted\ \ July\ temperature = 86.2$$

was found to be 347.6. So this line prediction is better (has a smaller total error) than the use of the single value, 86.8, to predict the July temperatures.

We just drew one line (found by trial and error) to predict the July temperature from the January precipitation. Is this the best line prediction we can find in the sense of smallest sum of squared residuals?

Actually no. It turns out that one can mathematically find the line that makes the sum of squared residuals as small as possible. This line is called the *least-squares line* since it minimizes (makes "least") the squares of the residuals.

There is a relatively simple formula for this line. The least-squares line has the equation

$$\hat{y} = a + b\,x$$

($\hat{y}$ denotes predicted value of y) where the slope *b* and the intercept *a* are given by

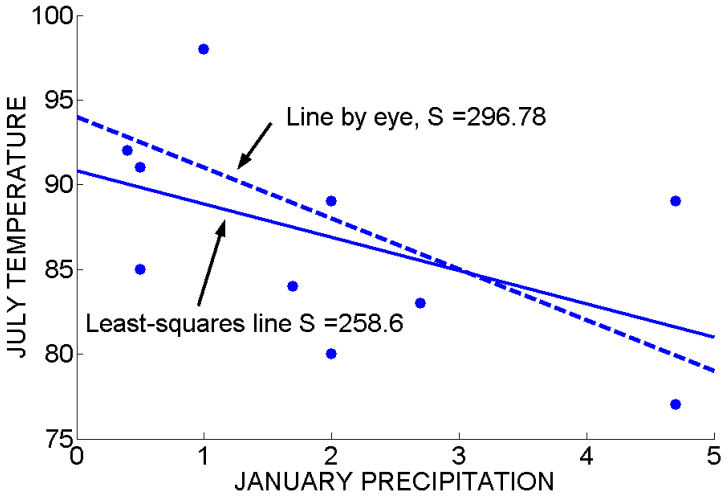
$$b = r\frac{s_y}{s_x},\ \ a = \bar{y} - b\,\bar{x}\,,$$

where r is the correlation coefficient, $\bar{x}, s_x$ are the mean and standard deviation of the x (horizontal) variable, and $\bar{y}, s_y$ are the mean and standard deviation of the y (vertical) variable. The formula for the intercept *a* reflects the fact that the least-squares line passes through the point ($\bar{x}, \bar{y}$).

In practice, we don't find the least-squares line by hand – instead, we use a computer program. In the below figure, the least-squares line

$$July\ temperature = -1.96 \times (January\ \ precipitation) + 90.8$$

has been drawn together with the line that we fit by eye. The sum of the squared residuals of this least-squares line is equal to 258.6; recall that the sum of squared residuals about our best line by eye was 296.78. As we would expect, this sum of squared residuals is smaller than the value we got from the line that we found by trial and error.



PRACTICE: WHAT IS LEAST-SQUARES?

The below table displays the average price of unleaded gasoline (in cents) in the United States for every four years from 1976 through 2004.

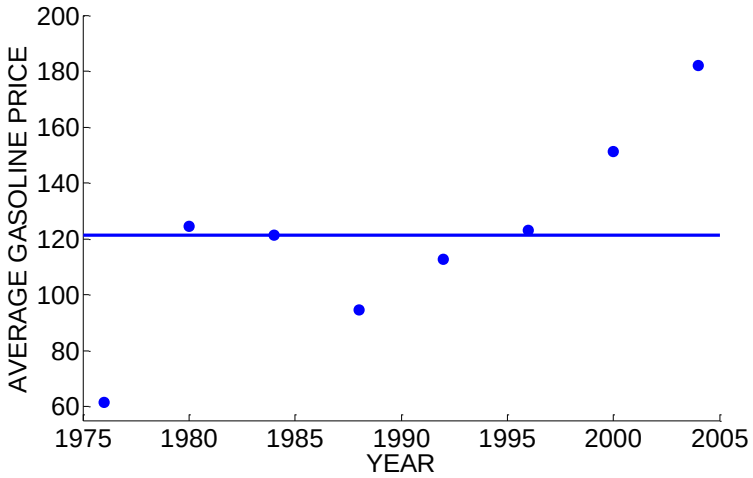
Year	price	predicted	residual	Residual squared
1976	61.4	121.3		
1980	124.5	121.3		
1984	121.2	121.3		
1988	94.6	121.3		
1992	112.7	121.3		
1996	123.1	121.3		

2000	151.1	121.3
2004	181.9	121.3
TOTAL		

1. Suppose we wish to predict the gasoline price using the mean value which can be computed to be 121.31 cents. The figure below shows a scatterplot of price against year with the basic prediction model

$$Average\ gasoline\ price = 121.31$$

drawn as a horizontal line on top of the scatterplot.



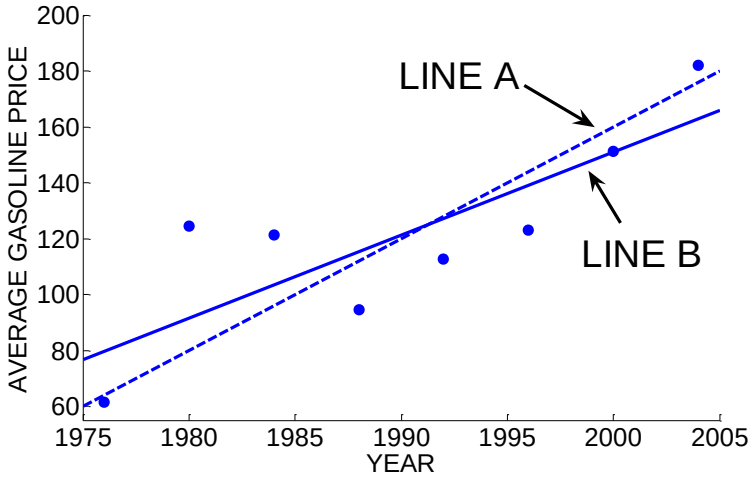
For each year in the table above, find the residual and the residual squared. Find the sum of squared residuals for this prediction of gasoline price.

2. Suppose instead that the price of gasoline is predicted using the "trial and error" formula

$$Average\ gasoline\ price = 120 + 4(Year - 1990).$$

(This is line A in the below figure.) In the below table, find the predicted prices for each year. Find the residuals, the squared residuals, and the sum of squared residuals.

Year	price	predicted	residual	residual squared
1976	61.4			
1980	124.5			
1984	121.2			
1988	94.6			
1992	112.7			
1996	123.1			
2000	151.1			
2004	181.9			
TOTAL				



3. The "least-squares" line was fit to these data – the equation of this line is

$$Average\ gasoline\ price = 121.31 + 2.98(Year - 1990).$$

which is Line B in the figure. As in part 2, find the predicted gas prices, the residuals, and the squared residuals using this line prediction. Find the sum of squared residuals.

4. To summarize your work, place the sum of squared residuals for each of the three fits in the below table. Which is the best fit, the next-best fit, and the worst fit in the sense of minimizing the sum of squared residuals? Explain why?

Fit	Sum of squared residuals
Constant value 121.31	
Line A: PRICE = 120 + 4(YEAR – 1990)	
Line B: PRICE = 121.31 + 2.98 (YEAR – 1990)	

Making appropriate and inappropriate predictions

The least-squares line allows us to make predictions about the response variable given a value of the explanatory variable. Let's return to the problem of predicting the July temperature of our hypothetical city Metropolis. Suppose we are told that this city's January precipitation is 4.0 inches. Then, using the least-squares line, we would predict the July temperature of Metropolis to average

$$\textit{July temperature} = -1.96 \times (4.0) + 90.8 = 82.96 \textit{ degrees.}$$

Suppose another city, say Emerald City, has a January precipitation of 10 inches. Using this line, we would predict the July temperature of Emerald City to average

$$\textit{July temperature} = -1.96 \times (10) + 90.8 = 71.2 \textit{ degrees.}$$

Thinking about this prediction, this seems that Emerald City is unusually cold in July. Looking back at the scatterplot, note that a January precipitation of 10 inches is off the graph – the highest precipitation value in our dataset was only 4.7 inches.

This illustrates one caution about the use of a best-line to make predictions. Based on the data, we are confident of a straight-line relationship *only* for the range of the values of the explanatory variables in the data. In this example, it only makes sense to predict July temperature for cities with January precipitation values between 0 and 4.7

inches. It would be inappropriate to use this line for values of January precipitations outside of this range.

PRACTICE: APPROPRIATE AND INAPPROPRIATE PREDICTIONS

Return to the earlier problem where we were examining the average gasoline prices of gasoline over a year. The least-squares fit was

$$\textit{Average gasoline price} = 121.31 + 2.98(\textit{Year} - 1990).$$

1. Would it be reasonable to use this line to predict the average gasoline price in 1980? Why or why not? If it is reasonable, make a prediction.

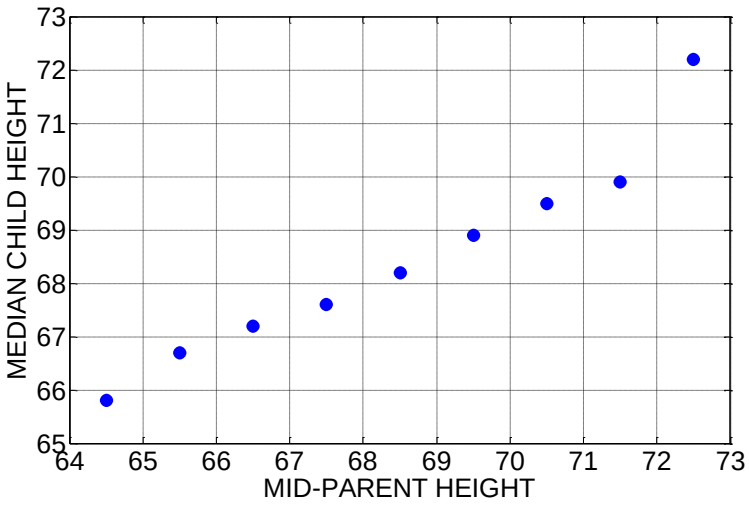
2. Would it be reasonable to use this line to predict the average gasoline price in 2050? Why or why not? If it is reasonable, make a prediction.

ACTIVITY: FITTING A LINE BY EYE TO GALTON'S DATA

DESCRIPTION: Francis Galton, in the famous 1886 paper "Regression Towards Mediocrity in Hereditary Stature," studied the degree to which children resembled their parents. For a large number of families, Galton measured the "mid-parent height" (the average of the heights of the mother and father), and the height of the child when fully grown (the "adult child height"). For each of the mid-parent heights 64.5 inches, 65.5 inches, and so on, the table below gives the median adult child height (in inches). (Galton multiplied all of the female children heights by 1.08, so that all of the children heights could be measured on the same scale.) A scatterplot of the two variables is shown to the right of the table.

MATERIALS NEEDED: A number of short pieces of spaghetti.

Mid-parent height	Median adult child height
64.5	65.8
65.5	66.7
66.5	67.2
67.5	67.6
68.5	68.2
69.5	68.9
70.5	69.5
71.5	69.9
72.5	72.2



1. Using the piece of spaghetti given to you by your instructor, fit a "good line" to the points on the scatterplot.

2. Find two points on your "good line." Write below your two points as ordered pairs.

	Mid-parent height	Median adult child height
Point 1		
Point 2		

3. Using your two points, find the slope and y-intercept of your line. (Show your work.) Write your equation of the line below.

MEDIAN CHILD HEIGHT = ☐ MID-PARENTS' HEIGHT + ☐

4. If the average of the mid-parents' heights is 66.5 inches, use your line to predict the median adult child's height.

5. From the table, the median adult child's height (when the mid-parent height is 66.5 inches) was 67.2 inches. Compute the residual.

6. To understand the difficulty of fitting lines by eye, collect for all of the students in the class the

- Computed slopes
- Computed y-intercepts

Graph the batch of slopes. Summarize this batch, including a description of the shape, typical value, and any outliers. Likewise, graph and summarize the batch of y-intercepts.

7. Can you explain why there is so much variation in the slopes and y-intercepts?

THE MEDIAN-MEDIAN LINE – A ROBUST ALTERNATIVE METHOD OF FITTING A LINE

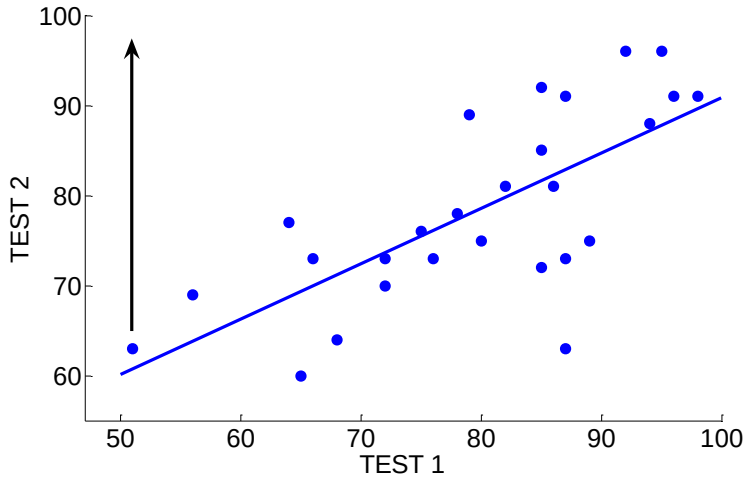
The least-squares line is the most common way of fitting a line to data. However, there is an undesirable characteristic of the least-squares method when there are outliers that are far away from the general pattern in the scatterplot. Here we demonstrate this problem and consider an alternative method of fitting a line that is less sensitive to outliers in the data.

Let's return to the dataset that contained two test scores for 27 students in a college class. (The dataset is shown below.) When we plot the Test 1 score against the Test 2 score, we see a positive association and we're interested in using a line to predict a student's Test 2 score from his Test 1 score. We fit a least-squares line that has the equation

Test 2 score = 29.44 + 0.614 Test 1 score.

Test1	Test2	Test1	Test2	Test1	Test2
82	81	95	96	78	78
79	89	98	91	89	75

92	96	87	73	85	72
96	91	80	75	75	76
72	70	76	73	56	69
86	81	72	73	87	91
66	73	94	88	51	63
87	63	64	77	68	64
85	92	65	60	85	85

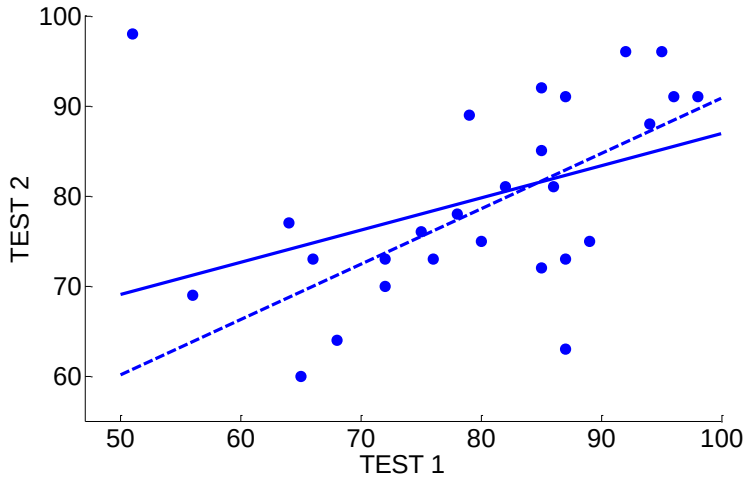


Let's introduce an outlier in the data. There is one student who received 51 on Test 1 and 63 on Test 2 – his scores are represented by the plotting point on the left side of the figure. Suppose that this student really got a 96 on Test 2 – with this change, the point on the left side is moved in the direction of the arrow. This student is really an outlier since his point is far away from the general group of points in the scatterplot.

Does the introduction of this outlier have any effect on our least-squares line? We recompute the line to be

Test 2 score = 51.22 + 0.357 Test 1 score

and graph this new line on the below scatterplot. Note that the least-squares fit has substantially changed – the single outlier seems to have flattened the least-squares line and the new line no longer seems to be a good fit to the general pattern of points. The problem with the least-squares line is similar to the problem of using a mean as an average of a batch of data in the presence of outliers. A least-squares line, like a mean, can be very sensitive to a few unusual data values that deviate from the main body of data.



A median-median line

There is another way of fitting a good line, the median-median line that is less sensitive to outliers. As the name suggests, this line is based on finding medians, rather than means, in regions of the scatterplot.

To compute this line, we first sort the data by the explanatory variable (here, Test 1) and divide the data into three equal-size groups (or nearly equal-size groups) by the ordered values of this variable. In this example, we have 27 points and we divide the data into the

- *Left group* – the 9 leftmost points
- *Middle group* – the 9 points in the center of the scatterplot
- *Right group* – the 9 rightmost points

The three groups of points are shown in the below table. Also in the below figure, we divide the points in the scatterplot by vertical lines, showing the left, middle, and right groups.

Left Group		Middle Group		Right Group	
Test1	Test2	Test1	Test2	Test1	Test2
51	63	76	73	87	63
56	69	78	78	87	73
64	77	79	89	87	91
65	60	80	75	89	75
66	73	82	81	92	96

	68	64	85	92	94	88
	72	70	85	72	95	96
	72	73	85	85	96	91
	75	76	86	81	98	91
Summary Point	66	70	82	81	92	91

For each group, we find a *summary point* that is the

(median of explanatory variable, median of response variable).

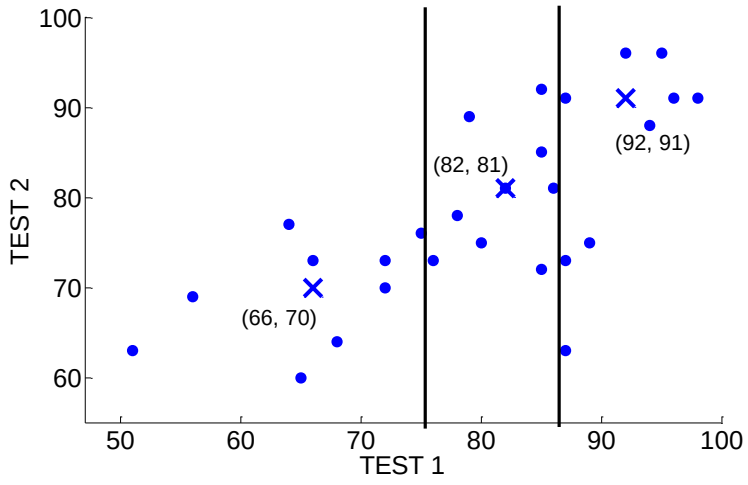
To illustrate, for the Left Group, the median of the explanatory variable Test 1, is

median{51, 56, 64, 65, 66, 68, 72, 72, 75} = 66

and the median of the response variable Test 2 is

median{63, 69, 77, 60, 73, 64, 70, 73, 76} = 70.

So the Left summary point is (66, 70). In a similar fashion, we find the summary points for the Middle and Right groups. We denote these points by (x_L, y_L), (x_M, y_M), and (x_R, y_R). They are shown in the table and plotted in the figure.



Now we can compute the median-median line:

1. The slope of the line is found by finding the slope between the Left and Right summary points.

$$b = \frac{y_R - y_L}{x_R - x_L}$$

Here the left and right summary points are (66, 70) and (92, 91) and so the slope is

$$b = \frac{91-70}{92-66} = 0.81.$$

2. Using each point, we can solve for the intercept by the usual formula $a = y - bx$. The intercept of the median-median line is the average of these three intercepts:

$$a = \frac{(y_L - bx_L) + (y_M - bx_M) + (y_R - bx_R)}{3}$$

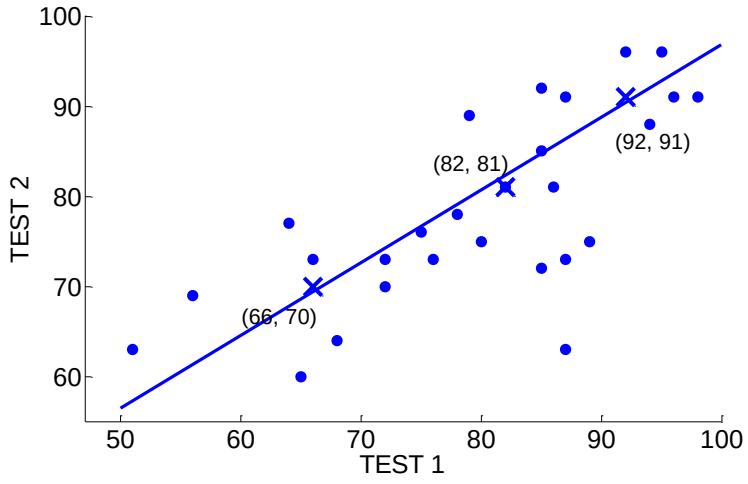
In this example, the intercept is computed to be

$$a = \frac{(70 - 0.81\times 66) + (81 - 0.81\times 82) + (91 - 0.81\times 92)}{3} = 16.05.$$

Summing up, our median-median line is

Test 2 score = 16.05 + 0.81 × (Test 1 score).

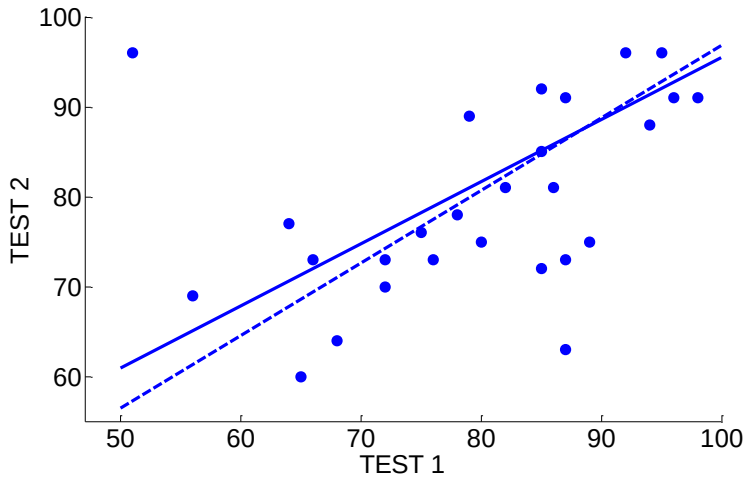
This line is displayed in the below scatterplot.



What was the point of introducing this new way of fitting a line? Remember that the least-squares line could be distorted by the introduction of a single outlying point. Suppose, as above, that we change the Test 2 grade of that one student from 63 to 96. The median-median line for this new dataset is

$$\text{Test 2 score} = 26.3 + 0.69 \times (\text{Test 1 score}).$$

We've plotted the original median-median line and the new one in the below figure. Note that the median-median line has changed, but this line is much less affected by the outlier than the least-squares line. To say it a little differently, the median-median is more robust or insensitive to outliers than is the least-squares line. In practice, both lines will give similar fits with no outliers present, but the lines can give very different results when there are outliers in the data.



PRACTICE: THE MEDIAN-MEDIAN LINE

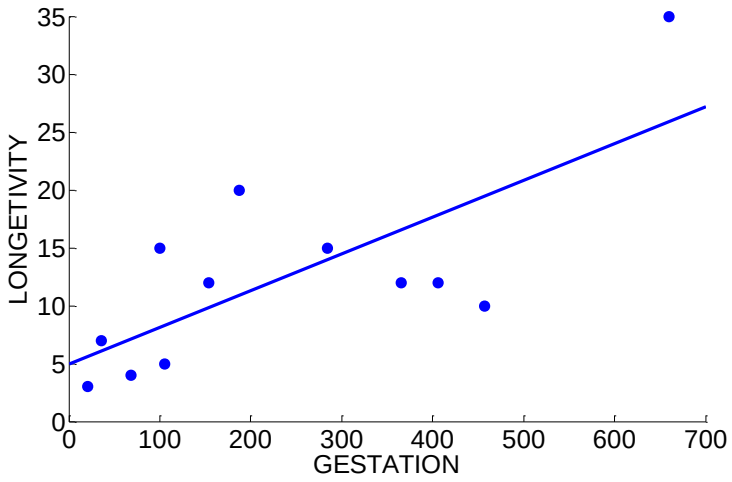
The table below gives the average gestation period (in days) and the average longevity (in years) for twelve animal species.

Animal	Gestation	Longevity	Summary points
Mouse	21	3	
Kangaroo	36	7	
Guinea pig	68	4	
Lion	100	15	
Beaver	105	5	
Sheep	154	12	
Baboon	187	20	
Cow	284	15	
Ass	365	12	
Camel	406	12	
Giraffe	457	10	
Elephant	660	35	

A scatterplot of the two variables is shown below. If we are interested in predicting an animal's longevity based on its gestation period, a least-squares fit is given by

LONGEVITY = .0318 GESTATION + 5.0.

This best line is drawn on the graph.



1. The dataset has been arranged by increasing values of the gestation period. Divide the dataset into three groups and find the summary point for each group.

2. Using the three summary points, find the equation of the median-median line.

3. Graph the median-median line on the scatterplot.

4. Are the least-squares and median-median lines similar in this example? If they are different, describe any special features of this dataset that might cause the two fits to be different.

TECHNOLOGY ACTIVITY – Fitting a "best line"

DESCRIPTION: In this Fathom activity, we will get some experience fitting lines to a scatterplot. Although there are several reasonably good straight-line fits, we will see the least-squares line is the one that is best in minimizing the sum of squared residuals.

1. Suppose you plan to fly from Detroit to St. Louis over Thanksgiving break. How much do you expect your round-trip plane fare to cost?
2. Suppose instead you decide to fly from Detroit to Seattle? How much do you think this will cost?
3. Why are your answers to 1. and 2. different?

4. Besides the fact that different plane trips cover different distances, why is there so much variation in plane fares? What other factors besides distance determine the size of a plane fare?

PART I: Fitting a line to remove the tilt in the scatterplot

Recently the author collected the lowest fares from Detroit to a number of U.S. cities. Also, I found the distance (in miles) of each city from Detroit. The data can be found in the Fathom document **airfares1.ftm**.

In this document, you will see

- A scatterplot of the distance (MILES) against the plane FARE.
- A line that passes through the point $(\bar{x}, \bar{y})$ and has slope equal to .7.
- A graph of the residuals of this particular line fit.

5. By playing with the slider, find a line that seems to "best fit" the points. (You find the line that removes the "tilt" in the residual plot.) Write the equation of your line below (put your slope in the box).

FARE = 283.1 + ☐ (MILES – 821.4)

6. For your line, find the residuals for the cities Chicago and Denver. (Look at the RESIDUAL column of the data table.) Verify (using your calculator) that these two residuals have been computed correctly.

7. Suppose you plan on flying from Detroit to Miami over the break. Predict what the airfare will be. (You can find a web site that gives the distance between cities.)

PART II: Understanding a "least-squares" line.

Open the Fathom document **airfares2.ftm**. You will see the same dataset.

8. Construct a scatterplot of MILES AND FARE.

9. Place a moveable line on the plot by selecting the scatterplot and then selecting the menu item Graph > Moveable Line.

10. By selecting the menu item Graph > Show Squares, you will see squares that correspond to the residuals from the line. You want to make the sum of the areas of the squares (the sum of squared residuals) as small as possible. Move the line and try to make the Sum of Squares as small as you can. Write down your final line and the sum of squares.

FINAL LINE: SUM OF SQUARES:

11. Now see how close your line is to the "least-squares line." (Select the menu item Graph > Least Squares Line.) Write down the least squares line and the sum of squares.

LEAST-SQUARES LINE: SUM OF SQUARES:

12. Compute the difference between the sum of squares for your line and the sum of squares of the least-squares line. This measures how well you did in part 10.

PLOTTING RESIDUALS

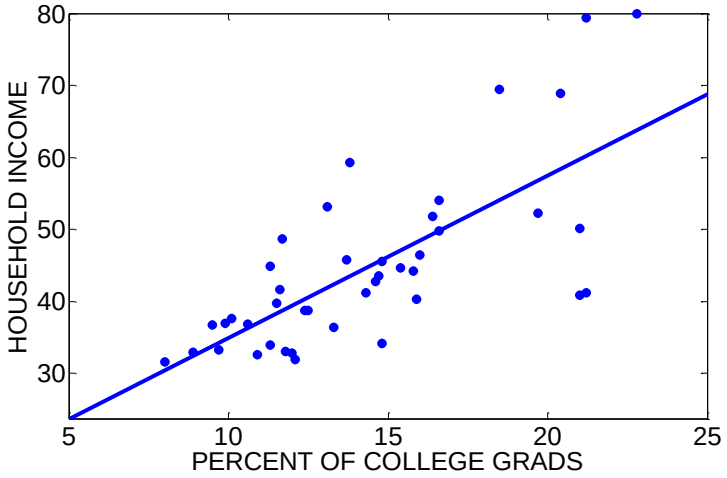
When we fit a line to a scatterplot relating two quantitative variables, we like to make the residuals small. The least-squares line is the one that makes the sum of squared residuals as small as possible. But the residuals, the differences between the observed and predicted responses, are important quantities in themselves. By graphing the residuals against the explanatory variable, we can see if our "best line" is a good description of the data.

In our cities dataset, the percentage of adults who completed four years of college and the average household income (in thousands of dollars) has been collected for 42 cities. Generally we think there is a relationship between these two variables: if a

greater percentage of adults are college educated, one might think that this would result in a higher average household income. Our objective here is to describe this relationship using a least-squares fit and then use a plot of the residuals to look for interesting points that deviate from the straight-line pattern.

We begin with a scatterplot of the two variables shown below. As expected, we see a positive relationship between the percentage of college graduates and household income. By use of a least-squares line, we get the relationship

HOUSEHOLD INCOME = 12.47 + 2.25 (PCT OF COLLEGE GRADS)



How can we interpret this least-squares line? The slope of the best line is b = 2.25 which means that if a city has a college graduation percentage that is 1% higher, then we would expect its average household income to increase by 2.25 thousand dollars. (Recall that a slope is the increase in the y-variable for a unit increase in the x-variable.)

Although the line is a general description of the relationship between college graduation percentage and household income, note that there are a number of data points that are far from the line. We can focus on these unusual points by plotting the residuals against the graduation percentage.

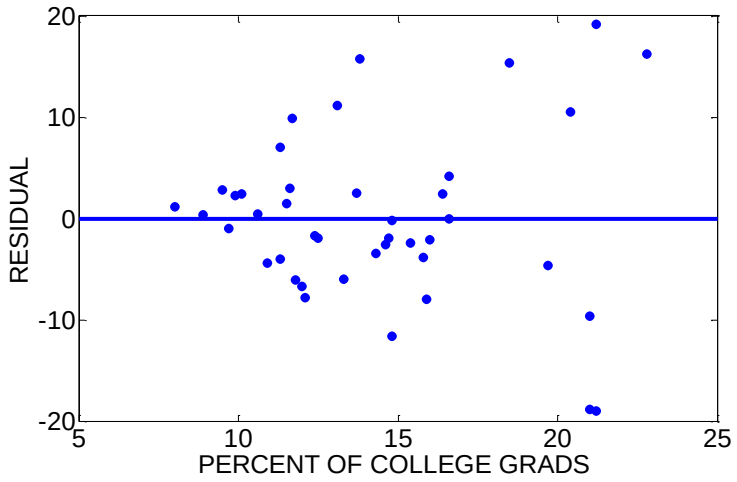
Let's review the computation of the residual. The first city Charlottesville has a college graduation percentage of 21 and an average household salary of $50.1 thousand. Using the line, we would predict this city's household income to be

PREDICTED INCOME = 12.47 + 2.25 x 21 = 59.7.

So the residual for Charlottesville would be

RESIDUAL = OBSERVED INCOME – PREDICTED INCOME = 50.1 – 59.7 = -9.6

Suppose we compute these residuals for all cities and graph the residuals against the explanatory variable (college graduation percentage). We get the following residual plot. (We typically add a horizontal line at the value RESIDUAL = 0 so it will be easy to spot the negative and positive residual values.)

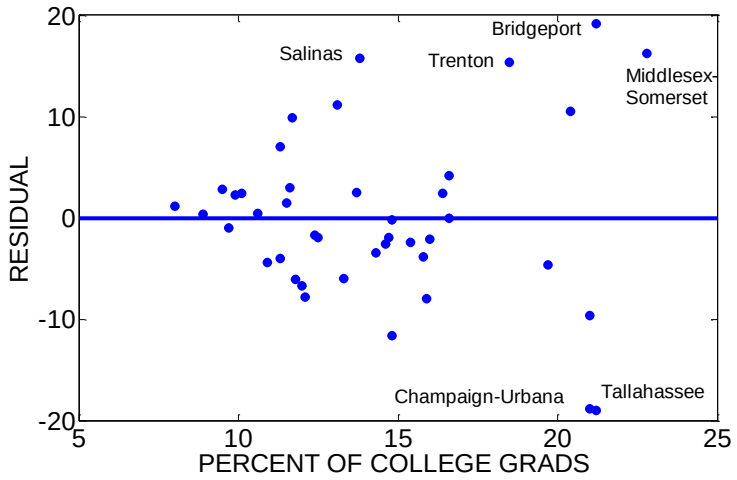


The residual plot focuses our attention on the deviations of the observed household incomes from the predicted values. When we look at a residual plot, we look for

- Systematic patterns of positive and negative residuals. For example, you might notice that there are many negative residuals on the left side of the plot and positive residuals on the right side of the plot. This indicates that the straight-line may not be a suitable fit to the data and an alternative method of describing the relationship may be necessary.
- Large positive and large negative residuals. In a residual plot, we look for large values that indicate points that are not close to the fitted line. We identify these

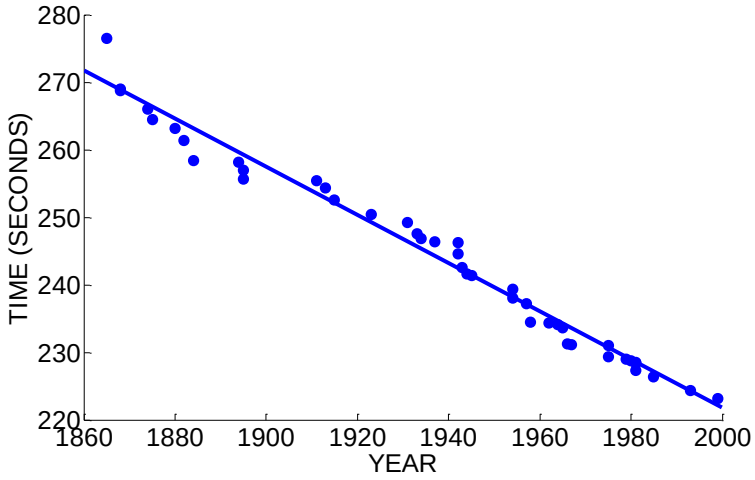
outlying points and think of any reason why these points don't follow the general pattern.

What do we see in our residual plot? There doesn't appear to be any systematic pattern of positive and negative residuals. Most of the residuals fall between –10 and 10 thousand dollars, but there are four large positive and two large negative residuals that seem to stand out. Looking back at our data, we see that the four large positive residuals correspond to Middlesex-Somerset (NJ), Bridgeport (CT), Salinas (CA), and Trenton (NJ). For these cities, their household incomes are much higher than one would predict on the basis of their college graduation rate. This is not surprising, since all four of these cities are located in expensive parts of the country (New Jersey, Connecticut, and California) where the cost of living is high. And the two large negative residuals correspond to Tallahassee (FL) and Champaign-Urbana (IL). These cities have household incomes that are smaller than one would predict on the basis of their college graduation rate. This is also not surprising since these are both college towns and the high college graduation rate reflects the college environment rather than the wealth of the general population. This discussion of unusually large residuals from the least-squares fit suggests that a city's average household income depends on more variables than just the education background of the population.



PRACTICE: PLOTTING RESIDUALS

Let's revisit the plot of the world record in the men's mile run that is graphed against the year where the record was obtained. Generally we see a straight-line pattern in the scatterplot – a least-squares fit is TIME = 933.94 - .36 YEAR .

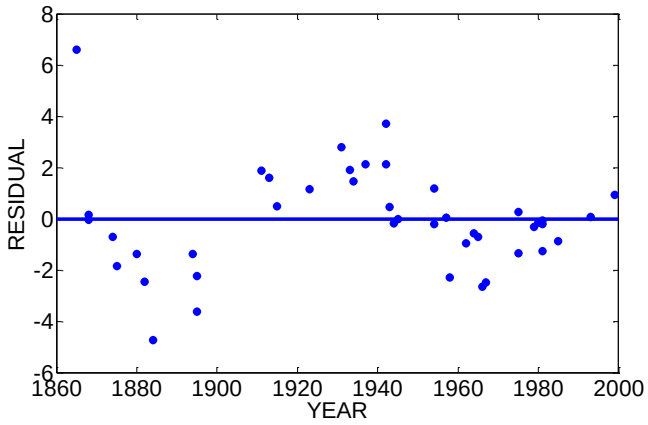


1. Based on the least-squares line, how much on average has the world record time decreased each year?

2. In the year 1895, the world record was 255.6 seconds. Use the least-squares equation to find the predicted time and compute the residual.

3. In the year 1937, the world record was 246.4 seconds. Compute the residual.

4. The residuals were computed for all points and a plot of the residuals against the year is displayed below.



Do you see any pattern in this residual plot? (Are there ranges of years where the residuals tend to be positive and when the residuals tend to be negative?) What does this say about the suitability of a straight-line relationship for the world record data?

5. Circle the point with the largest positive residual and the point with the largest negative residual. Is there any possible explanation for these large residuals?

6. As this book is written, the current world record for the mile run was set in 1999. Based on your work in this problem, does this surprise you? Explain.

TECHNOLOGY ACTIVITY: EXPLORING SOME OLYMPICS DATA

Open the Fathom file **summer_olympics.ftm**. This data gives the winning time in the men's 100, 200, 400, and 800 meter runs (in seconds) for each of the Summer Olympics from 1952 through 2004.

Focus on one race (either the 100m, 200m, 400m, or 800m) and see how the winning time for one race has changed from 1952 to 2004.

In your data exploration,

- Construct a scatterplot of the winning time (vertical) against the year (horizontal).
- Fit a good line to the points.
- Construct a residual plot.

Write a paragraph describing what you learned in this data analysis. In this paragraph, you should

- Give the equation of your best fit line and explain what it means. (How much on average is the winning time decreasing for each year?)
- Predict the winning time of the race in the next summer Olympics.
- Discuss any interesting features of the residual plot. Were there any particular years where the residual is unusually small or large? (Look for unusually small or large values in the residual plot.)

- The 1968 Olympics was unusual since it was held in Mexico City, a city at a high elevation, and many track and field records were set, such as the long jump. Do you notice anything unusual in your data in the 1968 time?

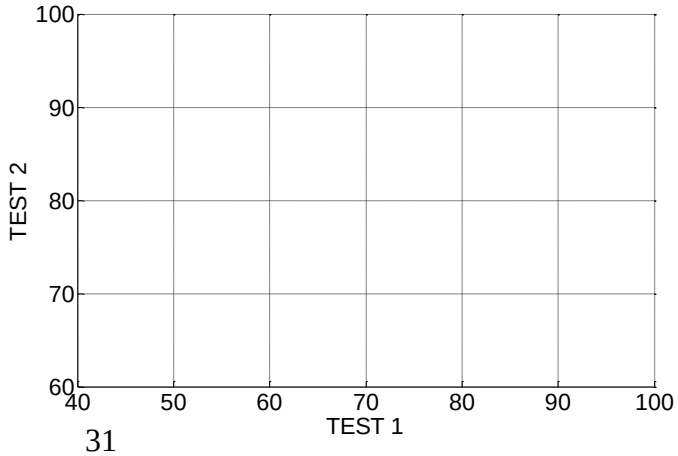
ACTIVITY: REGRESSION TO THE MEAN

DESCRIPTION: This activity demonstrates the "regression effect" that is generally unknown to many people. If you look up the word "regress" in the dictionary, it will tell you the word means to "go back." Suppose we collect two measurements from people that are positively correlated. We will see that there is a general tendency for a person's second measurement to go back, or regress to the mean.

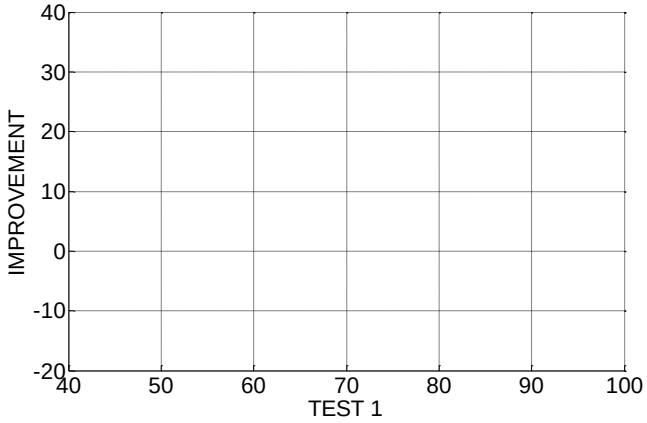
1. The following table gives the scores of 13 students on two tests in a statistics class. Construct a scatterplot of the Test 1 score (horizontal axis) against the Test 2 score on the below figure. Describe any relationship you see in the scatterplot. Is this relationship to be expected? Why?

STUDENT	TEST 1	TEST 2	IMPROVEMENT
1	96	87	
2	48	71	
3	75	80	
4	74	92	
5	88	97	
6	100	97	
7	51	77	
8	82	73	
9	80	87	
10	86	84	
11	76	64	
12	57	94	
13	56	79	

2. Next, compute the improvement TEST 2 – TEST 1 for each student and put the values in the table in the IMPROVEMENT column.



3. Construct a scatterplot of the improvement values (vertical axis) against the Test 1 scores (horizontal axis) on the axis below.



4. Do you see a pattern in this plot? Complete the following sentences. Students who had poor scores in Test 1 tended to _____ and students who did well in Test 1 tended to _____.

5. Some of you may have heard about the so-called "sophomore slump" in sports. This happens when a player does well in his/her rookie year and then slumps in the sophomore year. Some baseball people believe that this player may be struggling due to the pressure of maintaining the first-year performance. Based on what you have learned in this activity, is there an alternative explanation for the sophomore slump? Explain.

OPTIONAL ACTIVITY: For twelve players in a particular professional sport (such as baseball, basketball, or football), find their statistics for two consecutive seasons. Use this data to demonstrate or refute the regression effect.

DIFFERENT WAYS OF LOOKING AT RELATIONSHIPS

A typical human sleeps about eight hours a day. Is this a common amount of sleep among mammals? Or do humans sleep longer or shorter relative to other mammals?

To answer these questions, we need to obtain some relevant data. The table below gives the following characteristics of 24 types of animals:

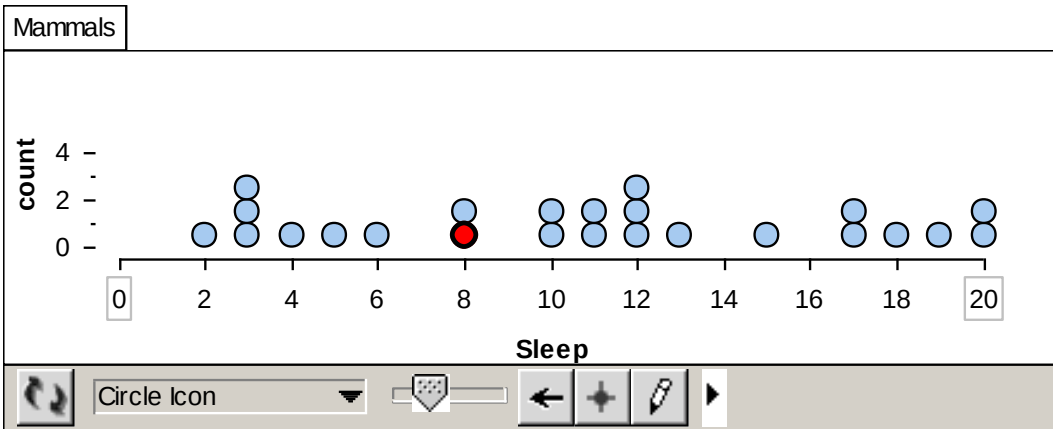
- MAMMAL -- the name of the animal
- SLEEP – the daily hours of sleep
- LIFESPAN – the life expectancy in years
- HEIGHT – the adult height in meters
- MASS – the adult mass in kilograms

Mammal	Sleep	LifeSpan	Height	Mass
African Elephant	3	70	4	6400
Asian Elephant	4	70	3	5000
Big Brown Bat	20	19	0.1	0.02
Bottlenose Dolphin	5	25	3.5	635
Cheetah	12	14	1.5	50
Chimpanzee	10	40	1.5	68
Domestic Cat	12	16	0.8	4.5
Donkey	3	40	1.2	187
Giraffe	2	25	5	1100
Gray Wolf	13	16	1.6	80
Grey Seal	6	30	2.1	275
Ground Squirrel	15	9	0.3	0.1
Horse	3	25	1.5	521
House Mouse	12	3	0.1	0.03
Human	8	80	1.9	80
Jaguar	11	20	1.8	115
Lion	20	15	2.5	250
N. American Opossum	19	5	0.5	5
Nine-Banded Armadillo	17	10	0.6	7
Owl Monkey	17	12	0.4	1

Pig	8	10	1	192
Rabbit	11	5	0.5	3
Red Fox	10	7	0.8	5
Spotted Hyena	18	25	0.9	70

Distribution of sleeping times

We focus on the relevant variable, SLEEP, that will help to answer our question. A dotplot of SLEEP for all 24 animals is shown below.



We see a lot of variation in the sleeping times -- some animals sleep, on average, 20 hours a day, and others sleep only 3 hours a day. Where does man stand in this sleeping time distribution? We label man's sleeping time with a black dot. We see that man's hours of sleep, 8, falls in the low end of this sleep distribution.

Since we observe a large spread of sleeping times it is natural next to try to explain why there is so much variation. Are there other variables in the dataset that might help in explaining the differences in sleep?

Let's consider LIFESPAN as a possible variable to help explain the variation in the response variable SLEEP. If we knew the lifetime of a particular animal, would this information be helpful in predicting its sleeping time?

Relating a categorical variable with a measurement variable

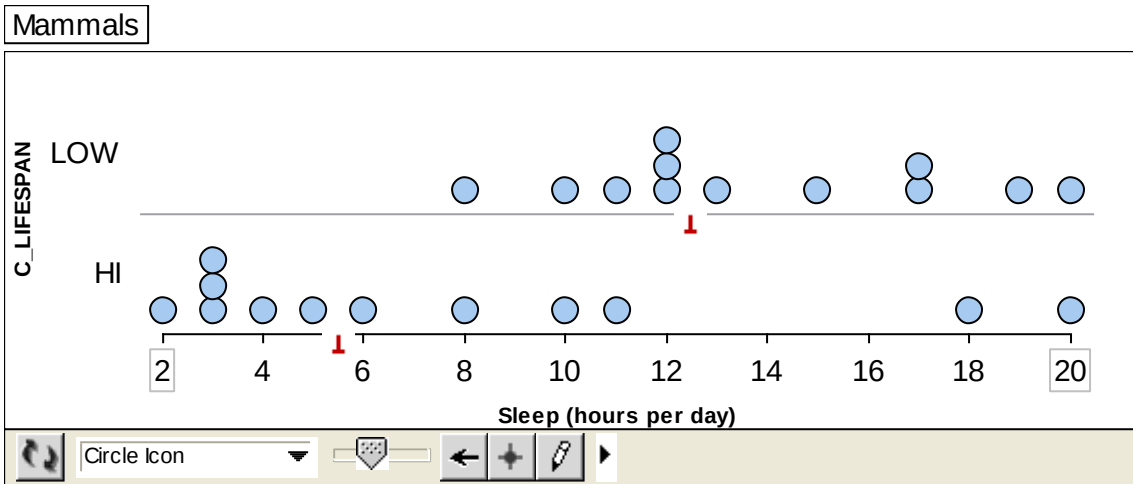
One way of studying the relationship between SLEEP and LIFESPAN is to divide the animal lifespans in two groups – the short-living animals and the long-living animals – and then compare the sleeping times of the two groups. This will be a familiar analysis for us, since we just described methods of comparing two batches in Topic D4.

We compute the median lifespan of the animals to be M = 17.5 years. We break the animals into two groups – the group living less than 17.5 years and the group living longer than 17.5 years – and then compare the sleeping times of the groups.

Here are the two groups of sleeping times:

SHORT-LIVING ANIMALS (LIFESPAN < 17.5 YEARS)		LONG-LIVING ANIMALS (LIFESPAN > 17.5 YEARS)	
Mammal	Sleep	Mammal	Sleep
Cheetah	12	African Elephant	3
Domestic Cat	12	Asian Elephant	4
Gray Wolf	13	Big Brown Bat	20
Ground Squirrel	15	Bottlenose Dolphin	5
House Mouse	12	Chimpanzee	10
Lion	20	Donkey	3
N. American Opossum	19	Giraffe	2
Nine-Banded Armadillo	17	Grey Seal	6
Owl Monkey	17	Horse	3
Pig	8	Human	8
Rabbit	11	Jaguar	11
Red Fox	10	Spotted Hyena	18

In the below figure, we plot parallel dotplots of the two groups of sleeping times; we have also computed the medians of each group and marked these values on the display. Reading from the graph, we see that the short-living animals sleep, on average, about 12.5 hours, and the long-living animals sleep on average about 5.5 hours. So there is indeed a relationship between lifespan and sleep – the short-livers tend to sleep about 7 hours a day longer than the long-livers.



Relating two categorical variables

The above analysis helps us to understand that sleeping time is indeed related to the lifespan of an animal. But there are other ways to describe this relationship and these alternative ways may be helpful in explaining this phenomenon to others or in understanding the relationship in more detail.

Suppose we categorize both the response variable SLEEP and the explanatory variable LIFESPAN into two groups. Above we defined short and long livers by the animals that lived shorter and longer than the median lifespan. Similarly, we can compute the median sleeping time M = 11 hours and define

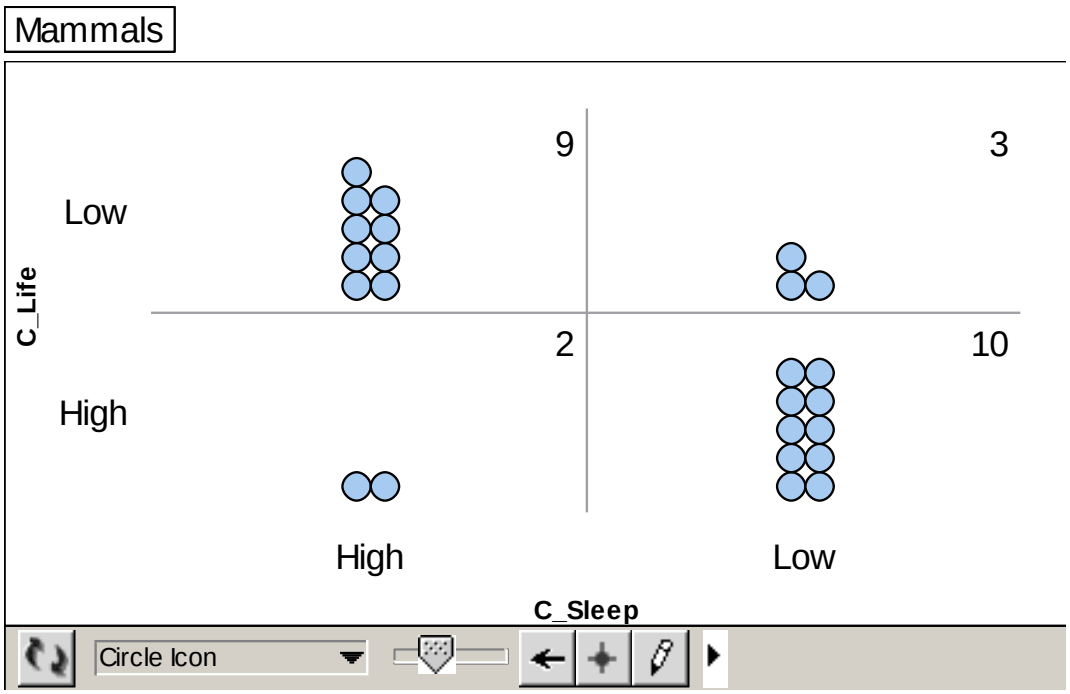
- Light-sleepers – animals that sleep 11 or fewer hours a day
- Heavy-sleepers – animals that sleep more than 11 hours a day

In the table below, we categorize all animals as HIGH or LOW on each of the two variables.

Mammal	Sleep	Type of Sleep	LifeSpan	Type of LifeSpan
African Elephant	3	LOW	70	HIGH
Asian Elephant	4	LOW	70	HIGH
Big Brown Bat	20	HIGH	19	HIGH

Bottlenose Dolphin	5	LOW	25	HIGH
Cheetah	12	HIGH	14	LOW
Chimpanzee	10	LOW	40	HIGH
Domestic Cat	12	HIGH	16	LOW
Donkey	3	LOW	40	HIGH
Giraffe	2	LOW	25	HIGH
Gray Wolf	13	HIGH	16	LOW
Grey Seal	6	LOW	30	HIGH
Ground Squirrel	15	HIGH	9	LOW
Horse	3	LOW	25	HIGH
House Mouse	12	HIGH	3	LOW
Human	8	LOW	80	HIGH
Jaguar	11	LOW	20	HIGH
Lion	20	HIGH	15	LOW
N. American Opossum	19	HIGH	5	LOW
Nine-Banded Armadillo	17	HIGH	10	LOW
Owl Monkey	17	HIGH	12	LOW
Pig	8	LOW	10	LOW
Rabbit	11	LOW	5	LOW
Red Fox	10	LOW	7	LOW
Spotted Hyena	18	HIGH	25	HIGH

Once we have categorized animals with respect to the two variables, we can divide the animals into four groups – the ones that are LOW on both variables, ones LOW on lifespan and HIGH on sleep, one HIGH on lifespan and LOW on sleep, and those that are LOW on both variables. In the following Tinkerplots display, each animal is represented by a dot, and the dots are divided into the four groups.



When both variables are categorical, then we can describe the relationship by the computation of percentages.

- Of the 12 short-living animals, we see from the figure that 9 or 9/12 = 75% are heavy-sleepers.
- In contrast, of the 12 long-living animals, we see that 2 or 2/12 = 16% are heavy sleepers.
- Since 75% is much higher than 16%, we can say that short-living animals are more likely to be heavy sleepers than long-living animals.

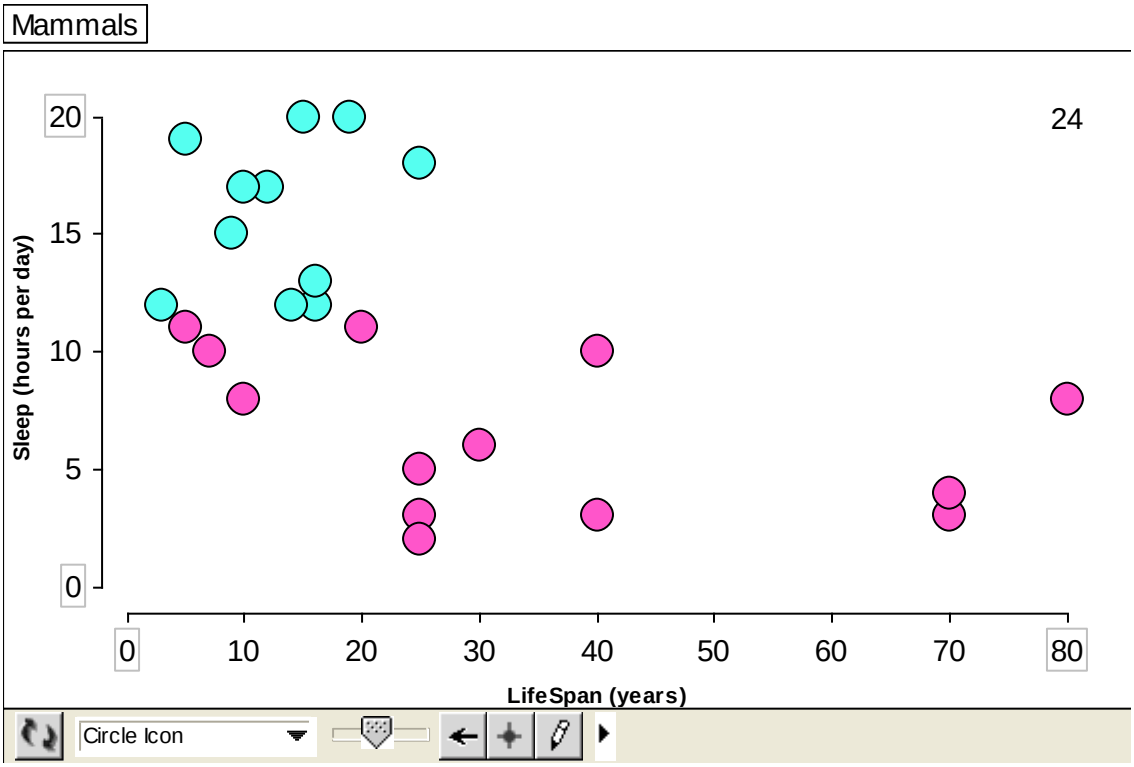
In Topic D7, we will focus on relationships between two variables that are both categorical.

Relating two measurement variables

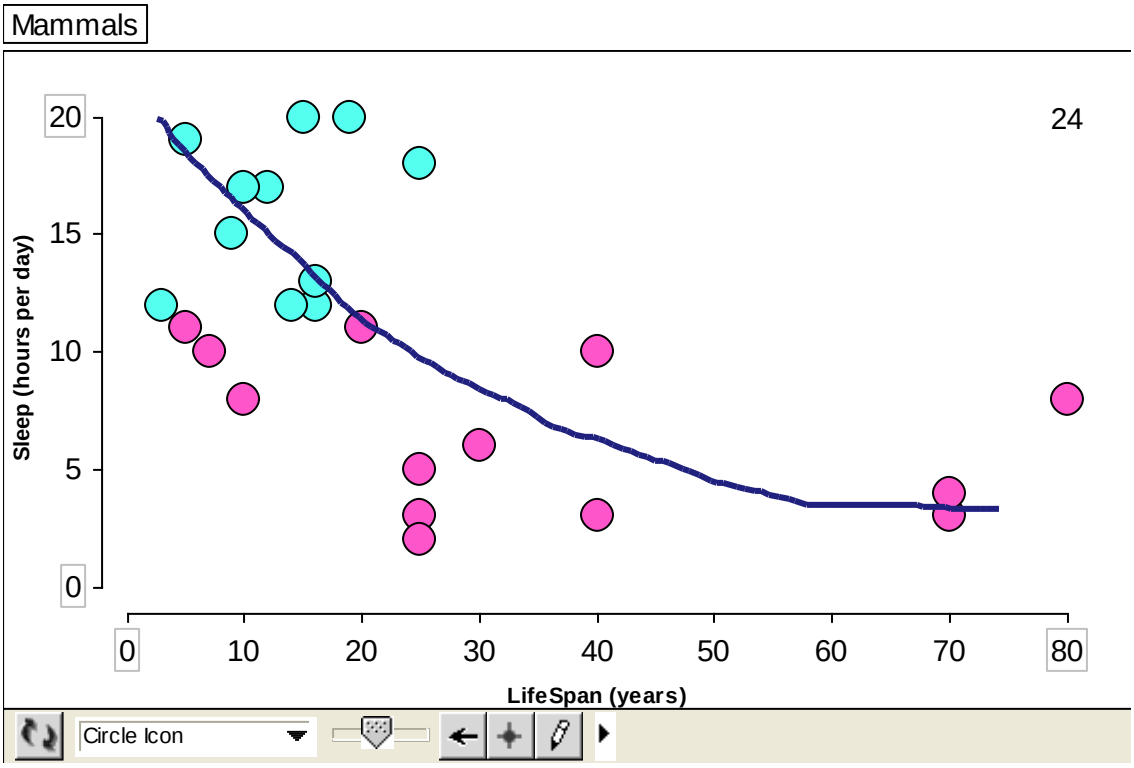
In both of the above analyses, we categorized animals as having short or long lifespans and compared the two groups of animals with respect to sleeping time. Is it possible to relate lifespan and sleeping time without this categorization?

When one has two measurement variables, a useful initial way of studying their relationship is by means of a scatterplot. Here one draws a grid of possible values of

sleeping time (vertical) and lifespan (horizontal) and the data values are represented by dots placed at the corresponding values of the two variables. We get the scatterplot display shown below.



We show a pattern in the scatterplot by drawing a curve through the points as we look at the display from left to right.



We see that the drawn curve has a negative trend – this means that as the animals' lifespans increase, the sleeping times decrease. In particular, we see animals that live only about 10 years tend to sleep about 15 hours and animals with lifespans of 70 years tend to sleep about 5 hours.

What is the best way of studying relationships?

We have illustrated three methods for understanding how an animal's sleeping time relates to its lifespan. Which is the best way for understanding this relationship?

Actually, there is not a "best" method in general. We use different descriptions depending on the problem on how we plan on communicating the relationship. The use of the scatterplot might seem like the best method because one loses information about sleeping time lifespan when one categorized them into HI/LOW groups. But it can be difficult to summarize the pattern of the relationship in a scatterplot and easier to describe the relationship when the variables are categorized and summarized as counts in a two-way table. For any method we use, it is important to be able to summarize how the

knowledge of the explanatory variables helps us predict the values of the response variable.

PRACTICE: DIFFERENT WAYS AT LOOKING AT RELATIONSHIPS

In the previous example, suppose we are interested in explaining the difference in sleeping times by the MASS variable (the weight of the animal in kilograms).

1. The median mass of the 24 animals is 70 kilograms. Divide the animals into two groups – the ones who weigh at most 70 kilograms and the ones who weigh more than 70 kilograms. In the below table, write the names and sleeping times for the animals in the two groups.

MASS AT MOST 70 KG	
Animal	Sleeping Time

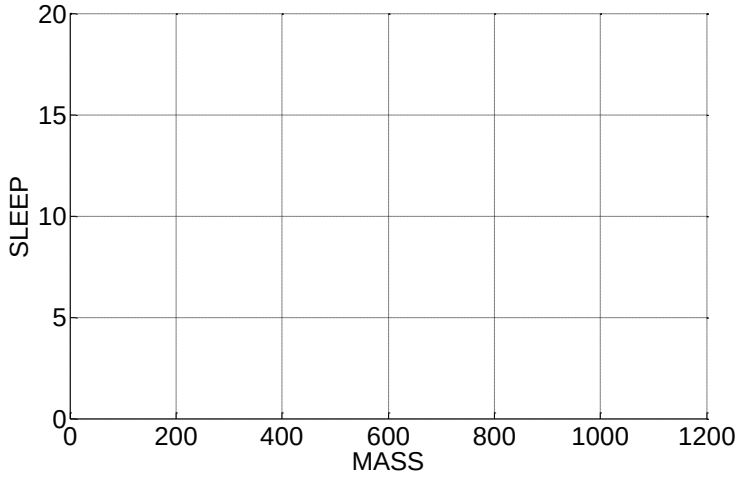
MASS MORE THAN 70 KG	
Animal	Sleep Time

2. By use of parallel boxplots, compare the sleeping times of the two groups of animals. Does one group of animals tend to sleep more on average than the other group? Explain.

3. Recall that the median sleeping time of the animals was 11 hours. As in the example, divide the animals into two groups – the light-sleepers (hours at most 11 hours) and the heavy-sleepers (hours exceeding 11 hours). Construct a two-way table of counts that categorizes the animals by mass and sleep. Using the data in this table, describe the relationship between mass and sleep.

		Sleeping Time	
		Low	High
MASS	Low		
	High		

4. As a final way of studying the relationship between mass and sleeping, construct a scatterplot of the two variables on the grid below. (To produce a reasonable looking display, we have limited the range of the horizontal axis from 0 to 1200 and two points will be off the graph.) Describe the general pattern in the scatterplot. As animals get heavier, do they tend to have lower or higher sleeping times?



TECHNOLOGY ACTIVITY – USING TINKERPLOTS TO STUDY RELATIONSHIPS

In this lab, you will use Tinkerplots to explore the relationship between a pair of variables.

1. First, you want to find two variables in a dataset where you suspect there is a relationship. A good source of interesting datasets can be found at DASL (the data and story library) at the website

`http://lib.stat.cmu.edu/DASL`

To find a interesting dataset where there is a relationship between a pair of variables, go to LIST ALL METHODS and find datasets that use the methods

Boxplot or Correlation or Regression or Contingency Table

2. BEFORE YOU GRAPH THE DATA, describe the dataset you will be using, discuss the variables you will be looking at, and what relationship you think you might find between these two variables. Why are you interested in this dataset?

[WRITE THIS DESCRIPTION ON THE TINKERPLOTS DOCUMENT.]

3. Copy the web address for the dataset that you are interested in on the Clipboard. Launch Tinkerplots. To get the data into Tinkerplots, go to

Import from URL from the File menu

and paste the web address of your dataset into the dialog box. Your data should be loaded into Tinkerplots.

4. Experiment with different graphs on Tinkerplots until you find one that is helpful for showing the relationship between the two variables.

5. Explain (AGAIN ON THE TINKERPLOTS DOCUMENT) what you have learned about the relationship between the two variables using the graph.

6. Repeat this work (steps 1 through 5) using two variables in a different dataset.

WRAP-UP

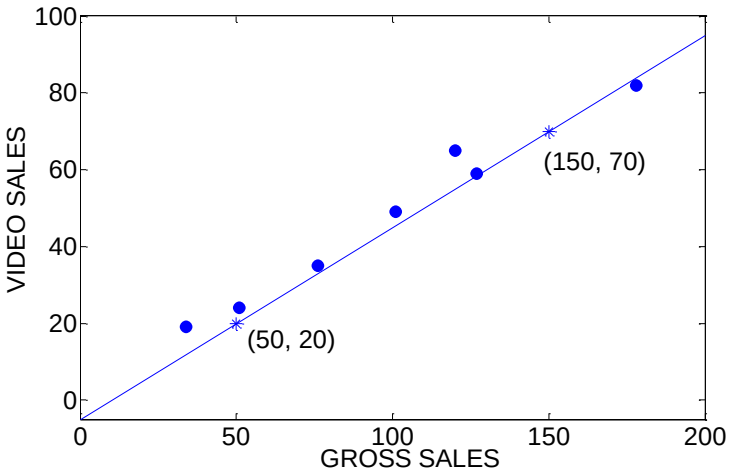
In this topic, we continued our study of relationships of two quantitative variables. The least-squares line is the standard method for fitting a straight-line to data. It is called least-squares since it is the line that makes the sum of squared residuals as small as possible. An alternative method of fitting a line, a median-median line, was described that is less sensitive to outliers than this least-squares method. When fitting a line, it is also helpful to construct a residual plot that focuses on the vertical distances of the points from the line. We look for patterns in the residual plot that indicate that a line may not be the best description of the relationship between the two variables. We demonstrated the regression effect where there is a general tendency with correlated data for the second observation to move back or regress towards the mean. We concluded by describing three approaches for studying the relationship between two quantitative variables. One approach for studying the relationship is to categorize each variable and make a comparison by comparing percentages as described in Topic D5. A second approach is to divide the explanatory variable into several categories and then compare the batches of values of the response variable using the methods described in Topic D4. A third strategy is to construct a *scatterplot* of the two variables and then use methods such as the computation of a correlation or the use of a best-line fit to understand the relationship.

EXERCISES

1. **Gross and Video Sales for Movies Starring Julia Roberts**

The table below lists the gross sales and video sales (both in millions of dollars as of March 2004) for seven movies featuring Julia Roberts. A scatterplot of these two variables follows.

MOVIE	GROSS SALES	VIDEO SALES
Conspiracy Theory (1997)	76	35
Dying Young (1991)	34	19
Hook (1991)	120	65
My Best Friend's Wedding (1997)	127	59
Pelican Brief, The (1993)	101	49
Pretty Woman (1990)	178	82
Something to Talk About (1995)	51	24



a. A line is fitted through the points (50, 20) and (150, 70), say. Find the equation of this line.

b. In the table below, compute the residuals using the fitted line that you found in (a).

Fitting the line from (a).

MOVIE	GROSS SALES	VIDEO SALES	Fitted value	Residual
Conspiracy Theory (1997)	76	35		
Dying Young (1991)	34	19		
Hook (1991)	120	65		
My Best Friend's Wedding (1997)	127	59		
Pelican Brief, The (1993)	101	49		
Pretty Woman (1990)	178	82		
Something to Talk About (1995)	51	24		

c. We can judge the goodness of the fit of this line by the sum of squared residuals. Compute the sum of squared residuals for the line you found in (a).

d. The "least-squares" line for these data is

$$\text{VIDEO SALES} = 2.39 + 0.46\ \text{GROSS SALES}$$

Compute the residuals for the least-squares line in the table below.

Fitting the least-squares line

MOVIE	GROSS	VIDEO	Fitted	Residual

	SALES	SALES	value
Conspiracy Theory (1997)	76	35	
Dying Young (1991)	34	19	
Hook (1991)	120	65	
My Best Friend's Wedding (1997)	127	59	
Pelican Brief, The (1993)	101	49	
Pretty Woman (1990)	178	82	
Something to Talk About (1995)	51	24	

e. Compute the sum of squared residuals for the least-squares line. Is this value larger or smaller than the sum of squared residuals for the line from part (a)? Are you surprised by this result?

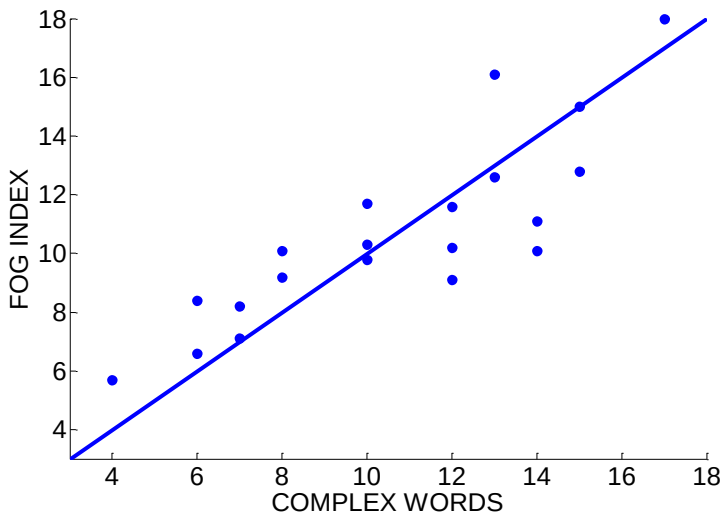
2. **Book Statistics**

The amazon.com website gives "text statistics" for many of the books it sells. For a particular book, the website displays the "fog index", the number of years of formal education required to read and understand a passage of text, and the "complex words", the percentage of words in the book with three or more syllables. The following table displays the complex words and fog index for a selection of 20 popular books.

Book	Complex Words	Fog Index	Residual
The Da Vinci Code	12	9.1	
Marley & Me	8	9.2	
The World is Flat	15	15	
Freakonomics	14	11.1	
Misquoting Jesus	13	16.1	
Power of Thinking Without Thinking	12	11.6	
The Mermaid Chair	7	8.2	
Memoirs of a Geisha	8	10.1	
The Five People You Meet in Heaven	6	6.6	
The Kite Runner	7	7.1	
In Cold Blood	10	9.8	

A Million Little Pieces	4	5.7
The Tipping Point	13	12.6
The Glass Castle	6	8.4
Collapse	17	18
Confessions of an Economics Hit Man	15	12.8
Curve Ball	14	10.1
A Mathematician at the Ballpark	12	10.2
Moneyball	10	10.3
Jim Cramer's Real Money	10	11.7

a. Suppose we are interested in predicting the Fog Index from Complex Words using the simple line formula FOG = COMPLEX. The graph below shows a scatterplot and this line. For each book, find the residual = actual Fog Index – predicted Fog Index and place the residuals in the table. Compute the sum of squared residuals.



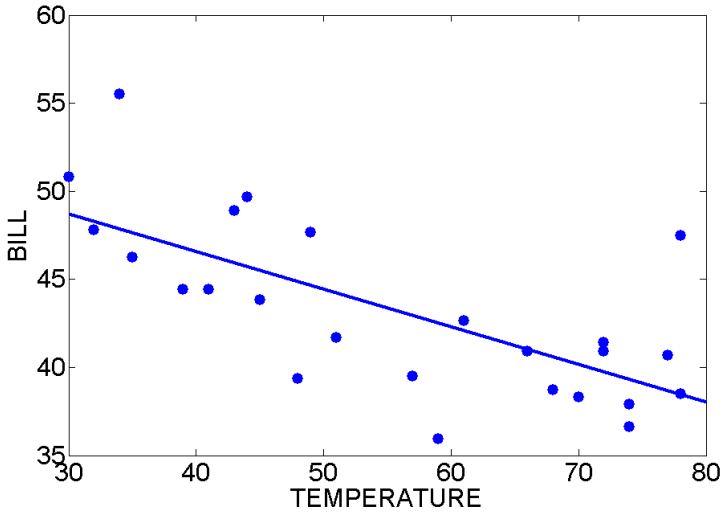
b. Find the book that has the residual of the largest size. Circle the point on the graph that has this largest residual.

c. The least-squares fit to these data is FOG = 2.91 + .73 COMPLEX. Find the residual for *The DaVinci Code* from this least-squares fit.

d. Suppose we compute the sum of squared residuals for the least-squares fit. Will this value be smaller or larger than the value of the sum of squared residuals you found in part a? Explain.

3. **Electricity Bills**

A homeowner collected the average temperature of a month (degrees Fahrenheit) and the amount of her electricity bill (in dollars) for that month. The figure below displays a scatterplot of temperature against the bill amount.



a. Does the scatterplot reveal a positive association between these variables, a negative association, or not much association at all? If there is an association, how strong is it?

b. Using the summary statistics shown below, determine the equation of the least squares (regression) line for predicting the electric bill from the average temperature. Record the equation of the line.

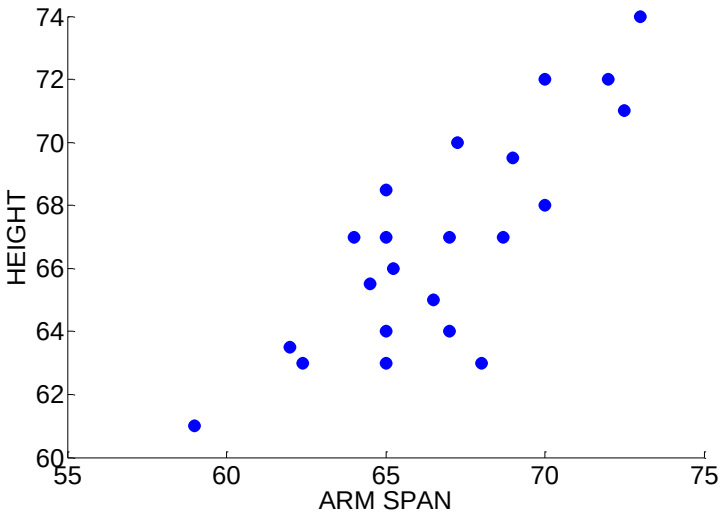
Variable	Temperature	Bill	Correlation
Mean	55.88	43.18	-0.695
Standard deviation	16.21	4.99	

c. In March of 1992, the temperature was 41 degrees and the electricity bill was $44.43. Use the equation you found in part b to determine (by hand) the fitted value and residual for March of 1992.

d. From looking at the graph, identify the points that have unusually large residual values. Were the electric bills higher or lower than expected for their average temperature?

4. **Height and Arm Span of Students**

It has been found that a person's arm span is strongly related to height. To investigate the nature of this relationship, arm spans and heights (both measured in inches) were measured for a class of 22 college students. A scatterplot of arm span and height is shown below.



a. The mean and standard deviation of arm span are given respectively by $\bar{x}$ = 66.73 inches and s_x = 3.52 inches, the summary values for height are $\bar{y}$ = 68.86 inches and s_y =3.48, and the correlation between the two variables is r = .813. Using these statistics, find the equation of the least-squares line predicting height from arm span.

b. Graph the equation of the least-squares line on the scatterplot.

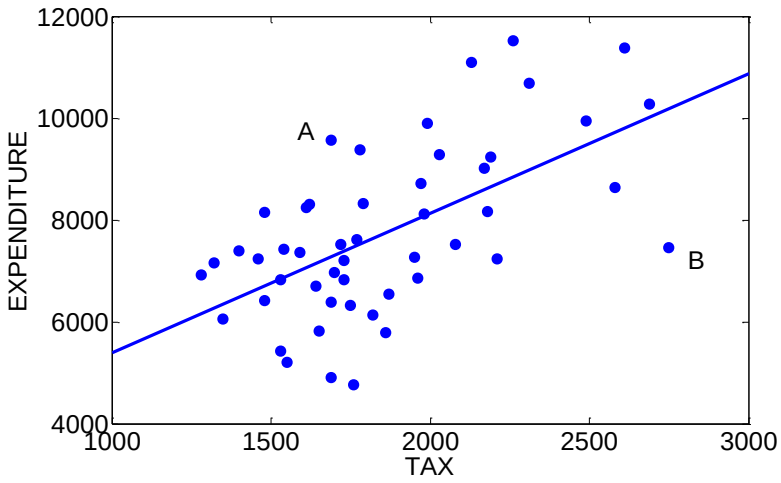
c. Suppose one student's arm span is 10 inches longer than another student's arm span. Predict how much taller the first student will be compared to the second student.

5. **Tax Revenue and Spending on Schools**

The 2004 Supplement of the *World Book Encyclopedia* gives the tax revenue per capita (the amount of tax collected per resident) and the public school expenditure per

pupil for each of the 50 states. Below a scatterplot of TAX vs. EXPENDITURE is displayed – a best-line (least-squares fit) is

EXPENDITURE = 2.74 TAX + 2650.



a. The slope of the best-fit line is _____ . This means that if a state decides to tax the residents an additional $100 per resident, one would predict that the expenditure per pupil would increase by _____.

b. The state of Ohio collected $1720 of tax per resident. Using the best-line equation, predict how much their expenditure should be per pupil.

c. Actually, it turns out that Ohio's expenditure per pupil was $7520. Compute the residual or the error in your prediction. (Remember that a RESIDUAL = ACTUAL y VALUE – PREDICTED y value.)

d. For North Dakota, the (TAX, EXPENDITURE) = (1760, 4770). Compute the residual.

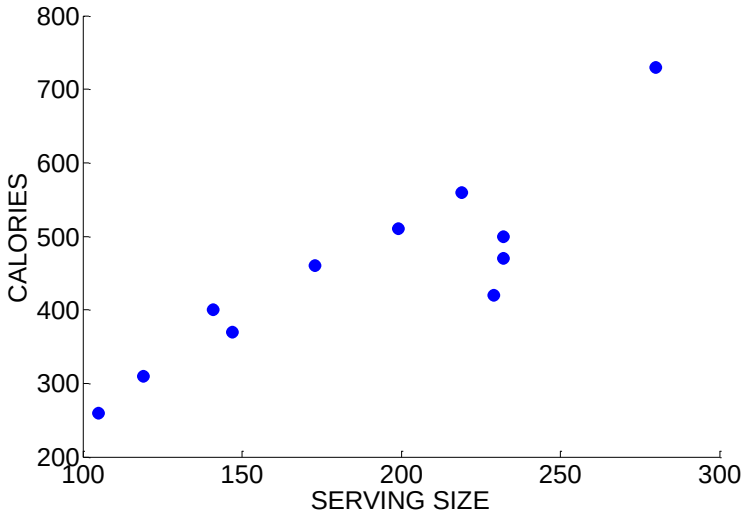
e. Two points are labeled in the graph. From the graph, compute (approximately) the value of the residual.

POINT "A" RESIDUAL: _____ POINT "B" RESIDUAL: _____

6. **Nutrition at a Fast-Food Restaurant**

McDonalds restaurant publishes nutritional information about all of the sandwiches they sell. The below table shows the serving size (in grams) and calories for a number of sandwiches. Suppose you are interested in understanding the relationship between the two variables. A scatterplot is shown below.

Sandwich	Serving size (gm)	Calories
Hamburger	105	260
Cheeseburger	119	310
Double Cheeseburger	173	460
Quarter Pounder® with Cheese+	199	510
Double Quarter Pounder® with Cheese++	280	730
Big Mac®	219	560
Big N' Tasty®	232	470
Filet-O-Fish®	141	400
McChicken ®	147	370
Premium Grilled Chicken Classic Sandwich	229	420
Premium Crispy Chicken Classic Sandwich	232	500



A least-squares line to these data is given by

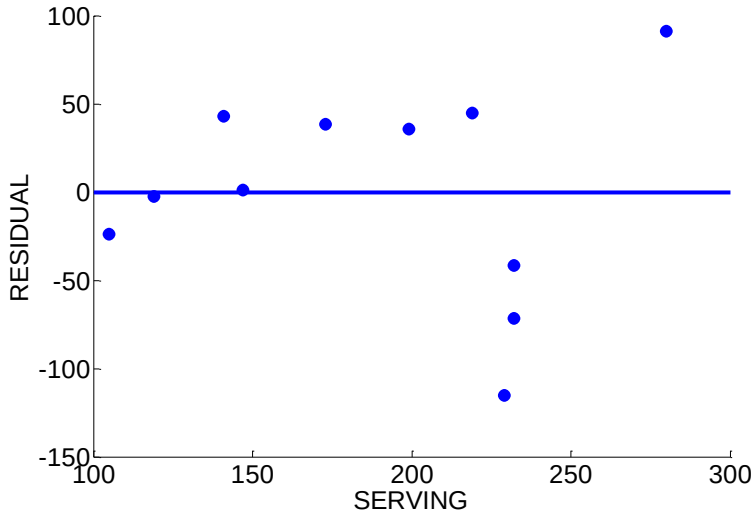
$$\text{CALORIES} = 71.0166 + 2.0274\ \text{SERVING\_SIZE}\ .$$

a. Suppose a sandwich's serving size is 200 grams. Predict the number of calories of this sandwich.

b. Suppose you are given the option to "super-size" a sandwich by making it 100 grams larger. How many extra calories will be in this super-size sandwich?

c. Compute the residual for Filet-O-Fish.

d. The residuals are computed for all sandwiches and a graph of the residuals against the serving size is displayed below. Are there any unusual points in this residual graph? Looking back at the data table, find the sandwiches that have these unusual residuals. Are these sandwiches different from the remaining sandwiches?



7. **High School Completion Rates**

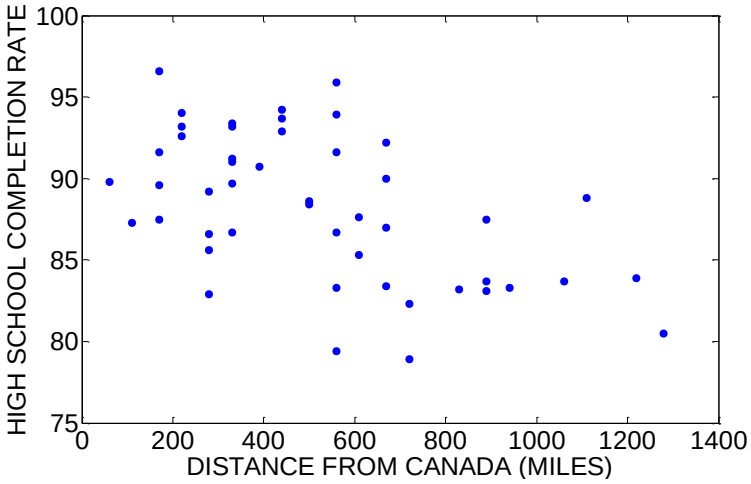
One measure of the educational level of the people who live in a particular state is the percent of adults who have received a high school diploma. The table below gives the adult high school completion rate (as a percentage) for each of the continental 48 states. (These data were obtained from the *1998 Wall Street Journal Almanac*.) Scanning this table, one notes considerable variability in these rates. For example, Nebraska (a northern state) has a high school completion rate of 95.9 %, while Georgia (a

southern state) has a rate of only 79%. That raises an interesting question. Is there a relationship between the state's high school completion rate and its geographic location? To help answer this question, the author got out his family's map of the United States and measured the distance from each state's capital to the Canadian border. These distances (in miles) are also recorded in the table.

State	Completion rate	Distance	State	Completion rate	Distance
Alabama	83.3	940	Nebraska	95.9	560
Arizona	83.7	1060	Nevada	83.4	670
Arkansas	87.5	890	New Hampshire	86.6	280
California	78.9	720	New Jersey	91	330
Colorado	87.6	610	New Mexico	83.7	890
Connecticut	92.6	220	New York	87.5	170
Delaware	93.7	440	North Carolina	85.3	610
Florida	83.2	830	North Dakota	96.6	170
Georgia	79.4	560	Ohio	89.6	170
Idaho	86.7	330	Oklahoma	83.1	890
Illinois	86.7	560	Oregon	82.9	280
Indiana	88.4	500	Pennsylvania	89.7	330
Iowa	94.2	440	Rhode Island	90.7	390
Kansas	92.2	670	South Carolina	87	670
Kentucky	83.3	560	South Dakota	93.2	330
Louisiana	83.9	1220	Tennessee	82.3	720
Maine	94	220	Texas	80.5	1280
Maryland	92.9	440	Utah	93.9	560

Massachusetts	91.2	330	Vermont	89.8	60
Michigan	89.2	280	Virginia	88.6	500
Minnesota	93.2	220	Washington	87.3	110
			West		
Mississippi	88.8	1110	Virginia	85.6	280
Missouri	90	670	Wisconsin	93.4	330
Montana	91.6	170	Wyoming	91.6	560

The figure below plots the distance from Canada (horizontal axis) against the completion rate (vertical axis).



a. By looking at the scatterplot, does there appear to be a relationship between distance and completion rate? What direction is the relationship? Is it a strong relationship?

b. Make an intelligent guess at the value of the correlation r.

c. Circle one point on the graph which corresponds to a state which is close to Canada and has a relatively small completion rate. Label this point A. Looking at the data, which state does this correspond to?

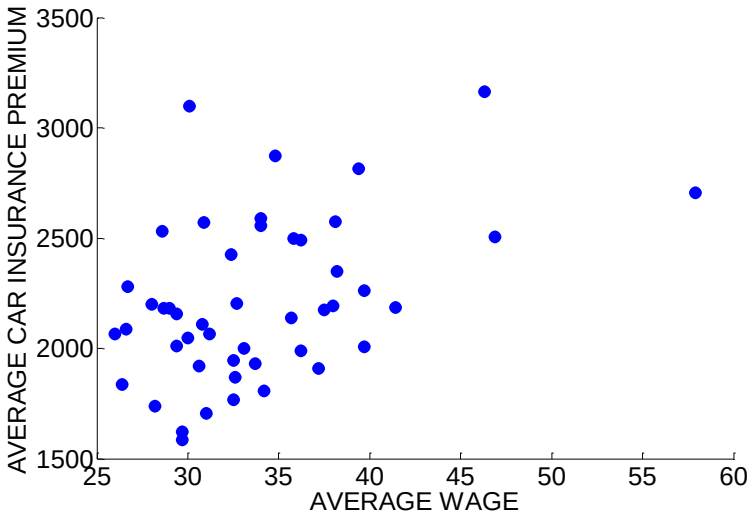
d. Circle a second point on the graph which corresponds to a state which is far from Canada and has a relatively large completion rate value. Label this point B. Which state does this point correspond to?

e. Can you think of another variable which is closely related to both completion rate and distance that might help explain the relationship that we observe in the scatterplot? (Such a variable is called a *lurking variable*.)

f. Suppose that a southern state is concerned about its relatively low high school completion rate. A state representative comments that maybe the solution to this problem is to move the residents of the state to a new location closer to Canada. Do you agree? Why or why not?

8. **Car Insurance Premiums and Average Salary**

One major expense in owning a car is insurance. There is a large variation in the cost of car insurance across states and is natural to wonder about the cause for this variation. For each state, two variables are recorded: the average annual car insurance premium in the year 2005 and the average wage in the year 2002. A scatterplot of these variables is shown below.



a. Is there a relationship between a state's average wage and its average car insurance premium? Describe the direction and strength of this association.

b. Make an intelligent guess at the value of the correlation r.

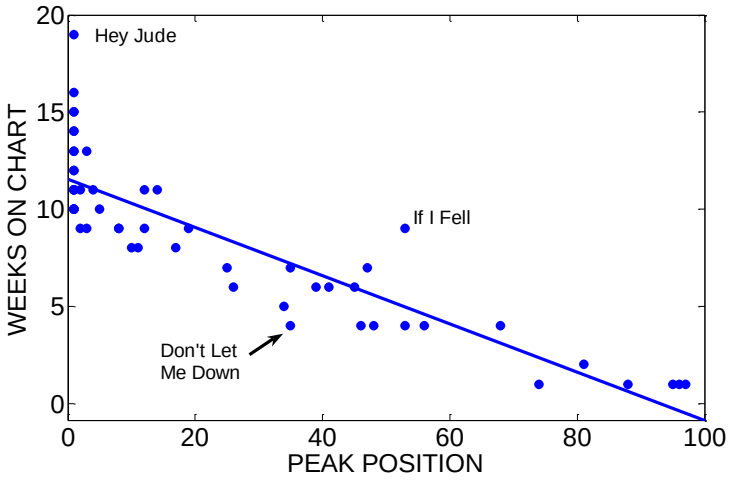
c. Would it be accurate to say that some state have high car insurance premiums since the workers in those states have higher incomes and therefore can afford these high premiums?

d. Is there a lurking variable present that might explain both the difference in wages and the difference in car insurance premiums? Explain.

9. **Beatles' Hit Songs**

The Beatles were a rock-and-roll band that achieved stardom in the 1960's. They recorded many albums that remain popular to the current day. Here we analyze characteristics of the entire set of singles that were released by the Beatles during their career. There were 58 Beatles singles that made the Billboard hit chart. The first song that reached number 1 on the chart was "I Want to Hold Your Hand" in 1964 and their last song to make number 1 was "Long and Winding Road" in 1970. For each Beatles single, two variables were recorded. The first variable which we call PEAK is the highest position on the Billboard hit chart, and the second variable WEEKS is the number of weeks that the song appeared on the Billboard Top 100 chart. One of the author's personal favorites, "Strawberry Fields," reached number 8 on the charts and stayed on the Top 100 for 9 weeks, so PEAK = 8 and WEEKS = 9.

The figure below displays a scatterplot of the PEAK and WEEKS variables for all 58 singles. To better understand the relationship between the two variables, we compute the least squares line which is given by WEEKS = 11.54 - 0.124 PEAK . This line is placed on top of the scatterplot in the figure.



a. Describe the general relationship between WEEKS and PEAK that you see in the scatterplot.

b. Suppose that a Beatles' song peaks at number 20 on the hit chart. Use the least squares line to predict how many weeks this song will stay on the Billboard Top 100.

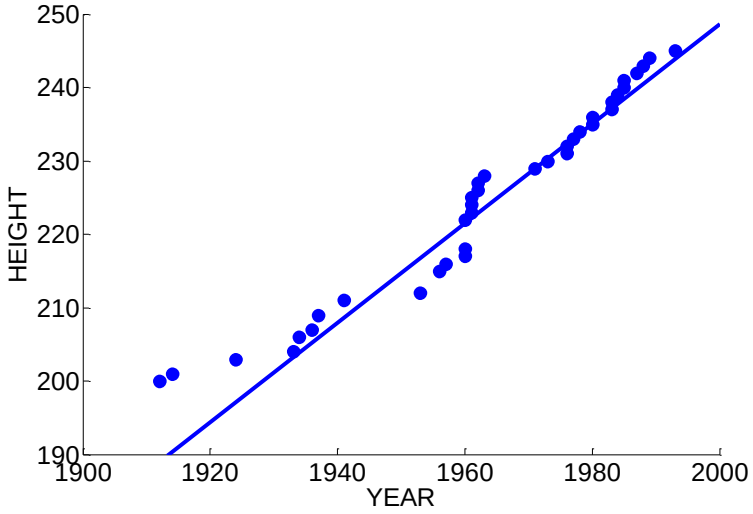
c. The point corresponding to the song ``Hey Jude" is labeled on the scatterplot. This song peaked at number 1 and stayed 19 weeks on the hit chart. Compute the residual for this song.

d. Two other songs, ``If I Fell" and ``Don't Let Me Down," are also labeled on the plot. By just looking at the plot, estimate the residuals for each of these songs. What is distinctive about these two songs that makes them have large residuals?

e. Where in the plot are the negative residuals (the points that fall below the line) located? Where are the positive residuals located? This pattern in the residuals suggests that a straight line is not the best fit to this particular data set.

10. **High Jump Record**

In most track and field events, there has been a general increase in the world record. To illustrate, the below figure graphs the word record achievement (in meters) in the men's high jump competition as a function of year.

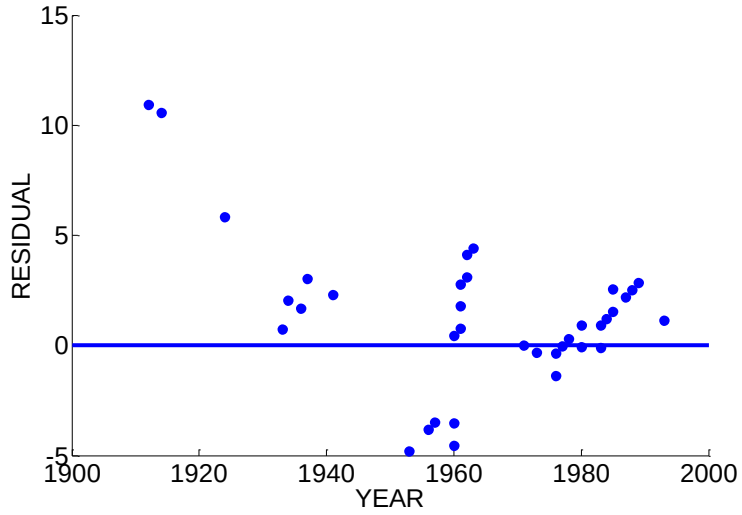


The pattern of achievement in the height of the record high jump can be described by the line HEIGHT = 0.6766 YEAR – 1104.6.

a. Generally, how much has the world record for the high jump changed for each year?

b. Can you use this line to predict the record in the high jump in the year 2050? Explain.

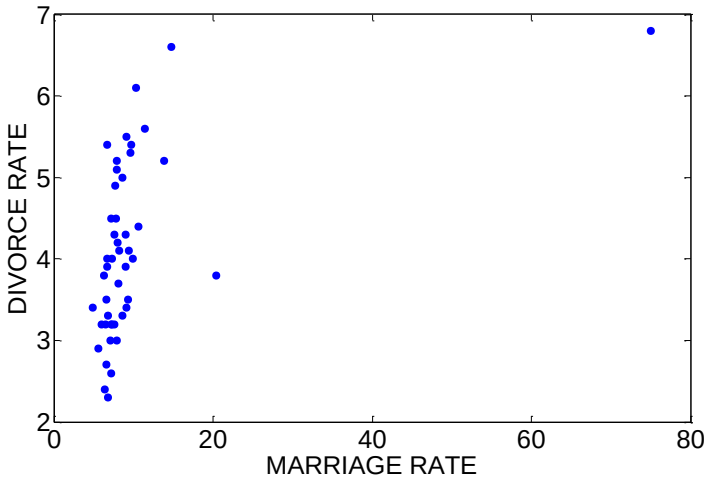
The below figure plots the residuals from this line fit as a function of year. This residual graph has a distinctive pattern.



c. There are four clusters of points in the residual graph corresponding to "before 1940", "1955 to 1960", "1960s", and "1970 to 2000". Describe the pattern of each cluster of points and explain what is says about the change in the high jump record.

11. **Marriage and Divorce Rates**

The *Statistical Abstract of the United Sates 2003* lists the marriage and divorce rates for the year 2001 of all of the states of the United States and the District of Columbia. To investigate a relationship between the two rates, we construct a scatterplot pictured below, where the divorce rate is plotted on the vertical axis and the marriage rate on the horizontal axis.



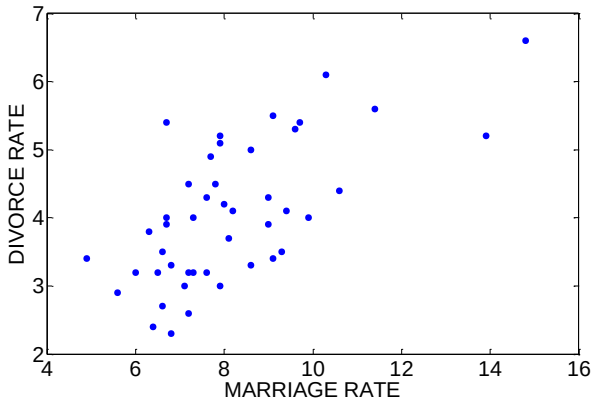
a. Describe in a few sentences any pattern or distinctive aspects of this scatterplot. Does a relationship appear to exist between the divorce rate and the marriage rate?

b. The correlation coefficient for this dataset is calculated to be r = .468. Explain in words what this value means.

c. Suppose that the marriage rates for Nevada and Hawaii (the unusual points in the scatterplot) are removed from the data. How would this change in the dataset affect the value of the correlation? Would the value of r go up, go down, or stay about the same?

d. Actually, the new value of r with Nevada and Hawaii removed is equal to r = .659. Why is this new value so different from the old value?

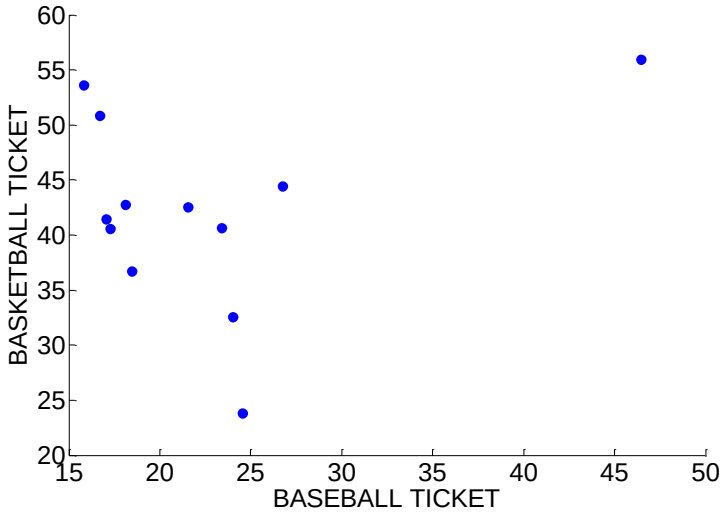
e. Suppose that we redraw the scatterplot with the Nevada and Hawaii points removed --- the new scatterplot is shown below. Does this change have an affect on the appearance of the scatterplot? Which scatterplot do you prefer --- the first one or the second one? Why?



12. **Cost of Sports Tickets**

There is variation in the cost of living among different cities in the United States. For example, it is more expensive to live in an eastern city, say Boston, than a city in the Midwest such as Minneapolis. Also one might think there is a relationship between the cost of two items for different cities. To investigate this, the below table displays the average cost of a ticket for professional baseball and professional basketball games for 12 cities. A scatterplot of these data is also drawn.

TEAM	MLB_Ticket	NBA_Ticket
BOSTON	46.46	55.93
SAN FRANCISCO	24.53	23.82
PHILADELPHIA	26.73	44.47
SEATTLE	24.01	32.54
TORONTO	23.4	40.67
DETROIT	18.48	36.75
CLEVELAND	21.54	42.52
FLORIDA	16.7	50.87
MINNESOTA	17.26	40.6
ATLANTA	17.07	41.43
TEXAS	15.81	53.6
MILWAUKEE	18.11	42.78



a. There is one outlying point in this graph. Identify the city that corresponds to this outlier.

b. Suppose the outlying point in the graph is removed. Describe the relationship in the graph and estimate the value of the correlation.

c. If one includes the outlying point, how do you think the value of the correlation will change? Why?

d. Use a statistics computer package or a calculator to compute the correlation for the full dataset and the dataset with the outlier removed. Is the value of the correlation sensitive to the inclusion of the outlying point? Explain.

13. **Gross and Video Sales for Movies Starring Julia Roberts**

In Exercise 1, the gross sales and video sales for seven Julia Roberts movies was considered.

a. Construct a median-median line predicting the video sales from the gross sales.

b. In Exercise 7, the least-squares line was given to be VIDEO SALES = 2.39 + 0.46 GROSS SALES. Compare the median-median and least-squares lines by finding the predicted values for each movie.

MOVIE	GROSS SALES	VIDEO SALES	Predicted value least-squares	Predicted value median-median
Conspiracy Theory (1997)	76	35		
Dying Young (1991)	34	19		
Hook (1991)	120	65		

My Best Friend's Wedding (1997)	127	59
Pelican Brief, The (1993)	101	49
Pretty Woman (1990)	178	82
Something to Talk About (1995)	51	24

c. Based on your work from part b, do the least-squares and median-median lines give similar predictions? Explain.

14. **Book Statistics**

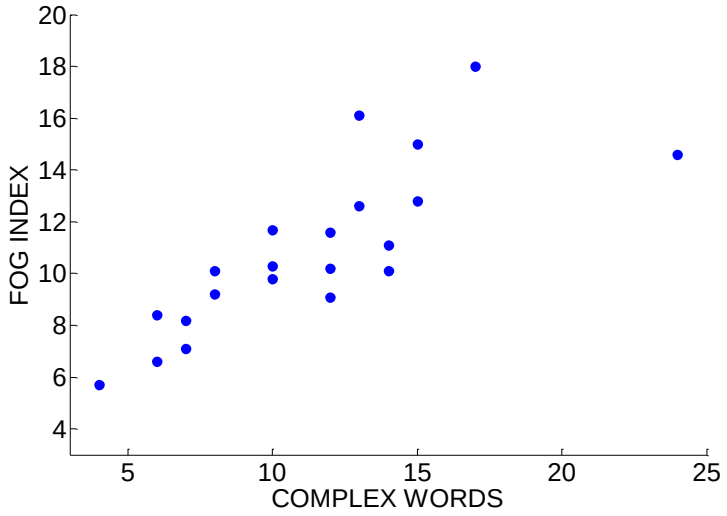
Suppose you are interested in predicting the Fog Index of a book from its Complex Words. (The data is given in Exercise 2.)

a. Find the median-median line by first finding the three summary points and then computing the slope and intercept.

b. The least squares fit to these data is FOG = 2.91 + .73 COMPLEX. Compare the median-median and least-squares fit by computing the predicted values for the following three books. Are the two lines similar with respect to their predictions for these books?

Book	Complex Words	Prediction of FOG from least-squares	Prediction of FOG from median-median
A Million Little Pieces	4		
The Glass Castle	6		
The Mermaid Chair	7		

c. Suppose that the book *Ordinal Data Modeling* with Complex Words = 24 and Fog Index = 14.6 is added to the dataset. A scatterplot of COMPLEX and FOG is drawn below – the new book is labeled on the graph. Recompute the median-median line for this new dataset. Is the median-median line sensitive to the addition of this new book? Explain.



d. The least squares fit to the original dataset was 2.91 + .73 COMPLEX and the least squares fit to the new dataset is FOG = 4.6 + .552 COMPLEX. Is the least squares line sensitive to the addition of this new book? Explain.

e. Based on your work from parts d and e, what is one advantage of the median-median line over the least-squares line?

15. **Batting Averages**

The table below gives the batting average (AVG) for 13 baseball players that played in the 2003 World Series between the New York Yankees and the Florida Marlins.

	2003 AVG	2002 AVG	Improvement
Jorge Posada	0.281	0.268	
Jason Giambi	0.250	0.314	
Alfonso Soriano	0.290	0.300	
Derek Jeter	0.324	0.297	
Bernie Williams	0.263	0.333	
Raul Mondesi	0.272	0.232	
Nick Johnson	0.284	0.243	
Ivan Rodriguez	0.297	0.314	

Derrek Lee	0.271	0.270
Luis Castillo	0.314	0.305
Mike Lowell	0.276	0.276
Juan Pierre	0.305	0.287
Juan Encarnacion	0.270	0.271

(a) Construct a scatterplot of the 2002 batting averages against the 2003 averages. Comment on any pattern in the plot.

(b) For each player, compute the improvement in batting average from the 2002 to the 2003 season (IMPROVEMENT = 2003 AVG – 2002 AVG). Place your values in the table.

(c) Construct a scatterplot of the improvement (vertical axis) against the 2002 AVG (horizontal axis). Describe the pattern of this pattern.

(d) Explain why this example illustrates the regression effect. Describe this concept in words that would be understandable by a layman.

(e) Suppose a baseball player has a great season and hits for a very high batting average (much higher than his previous seasons). Do you expect this player to have the same batting average the next season? Explain.

16. **Measuring School Achievement**

In many states, scores on standardized exams such as the ACT are used to rank high schools. The following table gives the average scores on an ACT exam for 14 Illinois high schools for the years 2001 and 2003.

High School	act2001	act2003Improvement
J STERLING MORTON WEST HIGH SCH	19.9	18.7
HAMPSHIRE HIGH SCHOOL	23.5	21.1
COULTERVILLE HIGH SCHOOL	22.6	18.7
VIRGINIA SR HIGH SCHOOL	19.7	19.1
MT ZION HIGH SCHOOL	22.7	21.4

BOGAN HIGH SCHOOL	17.1	16
BROWN COUNTY HIGH SCHOOL	21	19.4
ROCKFORD EAST HIGH SCHOOL	20.5	17.7
GENEVA COMMUNITY HIGH SCHOOL	23.7	21.5
CENTRAL HIGH SCHOOL	23.3	20.4
CARROLLTON HIGH SCHOOL	21.3	18.3
GLENBARD NORTH HIGH SCHOOL	22.7	20.8
VANDALIA COMMUNITY HIGH SCHOOL	21.5	19.2
ARCOLA HIGH SCHOOL	23	19.8

a. Construct a scatterplot of the 2001 ACT score and the 2003 ACT score. Describe the relationship you see in the scatterplot. Are you surprised by this relationship? Explain.
b. For each school, compute the improvement (2003 ACT score) – (2001 ACT score). Place the improvement values in the table.
c. Construct a scatterplot of the improvement values (vertical axis) against the 2001 ACT scores (horizontal axis). Describe the relationship you see in the graph.
d. Explain how this example illustrates regression to the mean.

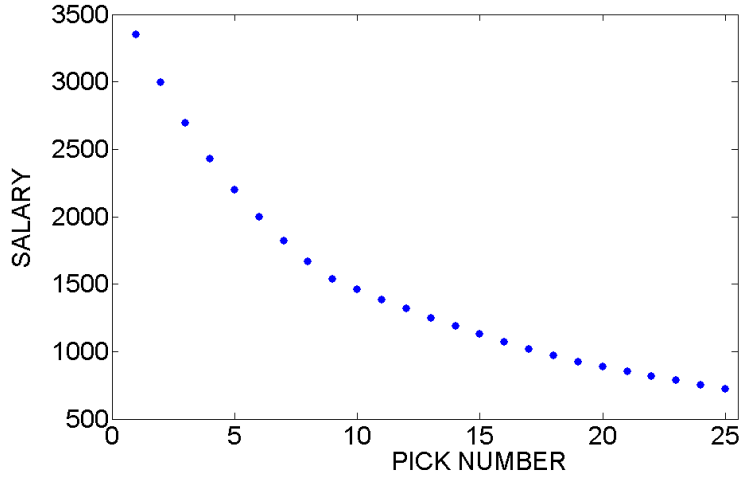
17. **Basketball Rookie Salaries**.

The table below pertains to basketball players selected in the first round of the 2003 National Basketball Association draft. It lists the draft number (the order in which the player was selected) of each player and the annual salary, in thousands of dollars, of the contract that the player signed. (Data is from the website InsideHoops.com.)

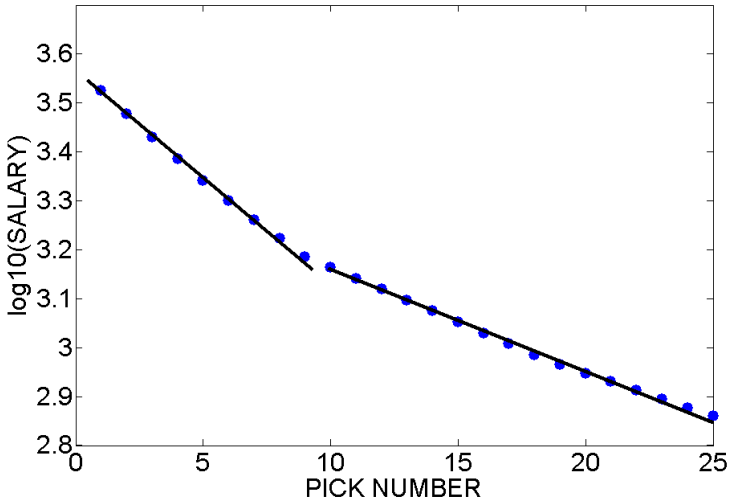
Pick Number	Salary	Pick Number	Salary	Pick Number	Salary
1	3349	**11**	1384	**21**	851
2	2997	**12**	1315	**22**	817
3	2691	**13**	1250	**23**	784
4	2426	**14**	1187	**24**	753
5	2197	**15**	1128	**25**	723
6	1996	**16**	1071		
7	1822	**17**	1018		
8	1669	**18**	967		

9	1534	**19**	923
10	1457	**20**	886

A scatterplot of salary against pick number is shown below.



Note that there is strong curvature in the plot and fitting a line will not provide a good description of the decreasing pattern. One can simplify the pattern by plotting the logarithm (base 10) of salary against pick number, obtaining the following scatterplot:



Note that this scatterplot can be fitted by two lines – one line for picks 1 through 9 and a second line for picks 10 through 25.

(a) Find the equation of the line fit to pick numbers 1 through 9. (The line will have the form log10(salary) = a + b (pick number).)

(b) Find the equation of the line fit to pick numbers 10 through 25.

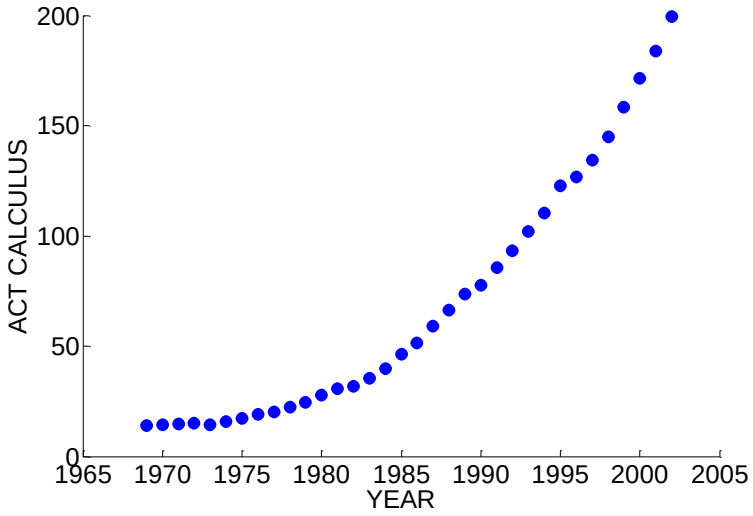
(c) By exponentiation of both sides of the equation, find the prediction equations for salary.

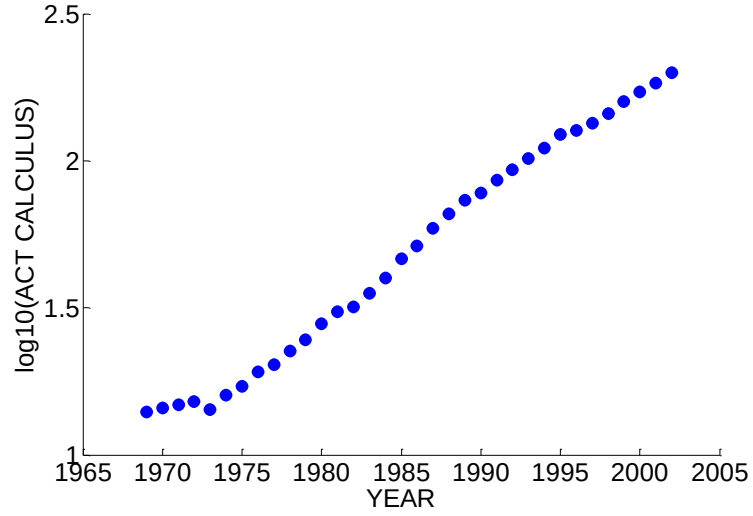
(d) What yearly salary would the regression line predict for the player drafted at number 5? How about for number 20?

(e) By how much does the regression line predict the salary to drop for each additional draft number? In other words, how much does a player stand to lose for each additional draft position which passes him by?

18. **Advanced Placement Tests**

The number of students who take the AP Calculus exam has increased since the introduction of this high school course in ???. The below graph displays the number of students taking this exam (ACT CALCULUS) for each year from 1969 to 2002. There is a strong association between YEAR and number of exam takers, but the association is not linear. If one transforms the y variable by a log (base 10) transformation, then the display ??? shows that the relationship between year and log(students) is approximately of the straight-line type.





a. Fit a median-median line to the (year, log10(students)) data.

b. If a line to these data has the form $\log(STUDENTS) = a + b\,YEAR$,

the model for STUDENTS has the form $STUDENTS = 10^{a+b\,YEAR} = 10^{a}(10^{b})^{YEAR}$. By using this form of the model, on average, what is the percentage increase in the number of students taking the AP calculus test each year?

19. **Brain Weight and Body Weight**

Are the brain weights of animals related to their body weights? In other words, does it require a larger brain to govern a heavier body? The brain weights (in grams) and the body weights (in kilograms) of 26 animals are given in the table below. Since there is

large variation in the data, it is convenient to record the log (base 10) brain weight and log body weight. Here you will explore the relationship between brain and body weight three ways.

a. Categorize each body weight as HIGH (larger than the median) or LOW (smaller than the median). Likewise, categorize each brain weight as HIGH (larger than the median) and LOW (smaller than the median). Construct a two-way count table. Using this table, describe the relationship between brain weight and body weight.

b. Consider the log brain weights of the LOW body weight animals, and the log brain weights of the HIGH body weight animals. By using 5-number summaries and parallel boxplots, compare the two groups of log brain weights.

c. Construct a scatterplot of the log body weights (horizontal axis) and the log brain weights (vertical axis). Describe the general pattern in the scatterplot. Are there any points in the scatterplot that seem to be different from the general pattern? (Which animals do these correspond to?)

d. Which way (a, b, or c) do you think is best for describing the relationship between brain and body weights? Why?

Species	Body wt (kg)	Log_body_wt	Brain wt (g)	Log_brain_wt
Mountain beaver	1.35	0.13	8.1	0.91
Cow	465	2.67	423	2.63
Gray wolf	36.33	1.56	119.5	2.08
Goat	27.66	1.44	115	2.06
Guinea pig	1.04	0.02	5.5	0.74
Diplodocus	11700	4.07	50	1.7
Asian elephant	2547	3.41	4603	3.66
Donkey	187.1	2.27	419	2.62
Horse	521	2.72	655	2.82
Polar monkey	10	1	115	2.06
Cat	3.3	0.52	25.6	1.41
Giraffe	529	2.72	680	2.83

Human	62	1.79	1320	3.12
African elephant	6654	3.82	5712	3.76
Triceratops	9400	3.97	70	1.85
Rhesus monkey	6.8	0.83	179	2.25
Kangaroo	35	1.54	56	1.75
Hamster	0.12	-0.92	1	0
Mouse	0.023	-1.64	0.4	-0.4
Rabbit	25	1.4	12.1	1.08
Sheep	55.5	1.74	175	2.24
Chimpanzee	52.16	1.72	440	2.64
Brachiosaurus	87000	4.94	154.5	2.19
Rat	0.28	-0.55	1.9	0.28
Mole	0.122	-0.91	3	0.48
Pig	192	2.28	180	2.26

20. **Lengths of Movies**

The table below gives the title, year made, and length (in minutes) for 40 movies randomly selected from the *Leonard Maltin's Movie and Video Guide* (1996). One is interested if there is a relationship between the year the movie is made and its running length.

a. Suppose the year of the movie is categorized as "old", made before 1960, and "new" made in the year 1960 or later. By use of five-number summaries and parallel boxplots, compare the times of the old and new movies.

b. Suppose you also categorize the movies as "short", 90 minutes or shorter, and "long", over 90 minutes. Construct a two-way table categorizing movies by year (old or new) and length (short or long). Based on this table, describe the relationship between year and length.

c. Construct a scatterplot of year (horizontal scale) against length (vertical scale). Describe the general pattern of the scatterplot.

Title	Year	Length	Title	Year	Length

The Twinkle in God's Eye	1955	73	Sleep My Love	1948	97
Dakota	1988	97	City Lights	1985	85
Evergreen	1934	90	Hambone and Hillie	1984	89
The Raven	1963	86	The Great Waldo Pepper	1975	107
Hitler--Dead or Alive	1943	70	You Only Live Twice	1967	116
The Ravine	1969	97	The Unholy Three	1930	72
Hold Back Tomorrow	1955	75	The Boogens	1981	95
Lady Dracula	1973	80	Jason's Lyric	1994	119
Kronos	1957	78	Divided Heart	1954	89
Descr A Ticklish Affair	1963	89	The Cockeyed Miracle	1946	81
She Demons	1958	80	The Siege at Red River	1954	81
Okinawa	1952	67	The Stone Boy	1984	93
Bachelor Apartment	1931	77	The Mutineers	1949	60
The Romantic Age	1949	86	Flash and the Firecat	1975	84
Valley of the Dragons	1961	79	The Amazing Transparent Man	1960	58
The Miracle Worker	1962	107	Night of the Dark Shadows	1971	97
Shout at the Devil	1976	119	Windwalker	1980	108
Our Man in Havana	1960	107	House Party 3	1994	94
Falcon Strikes Back	1943	66	Blondie Has Servant Trouble	1940	70
Smash Up: The Story of a Woman	1947	103	Seminole Uprising	1955	74