Chapter 1

Test

Complicated physical and social phenomena rarely behave linearly, but sometimes data points lie close to a straight line. When that happens you can use a spreadsheet to construct a linear approximation. Sometimes that's useful and informative. Sometimes it's misleading. Common sense can help you understand which.

Chapter goals:

Goal 1.1. Draw regression lines using Excel. Interpret regression lines.

Goal 1.2. Recognize when rounding too much distorts conclusions.

Goal 1.3. Think about causation vs correlation.

1.1 Climate change

Climate change (global warming) is a current hot topic. How rapidly is the Earth's average temperature increasing? What might the consequences be? What is the cause? What might we do about it? Should we try? The science is complex and the politics even more so. In a course like this we can't begin to unravel those complexities. But for just a taste of the analysis, we will briefly look at some data on the average temperature of the Earth and the concentration of carbon dioxide (CO_2) in the atmosphere in recent history. The spreadsheet `www.cs.umb.edu/~eb/qrbook/./EarthData.xlsx` has data we downloaded from `www.earth-policy.org/data_center/C23`.

The chart on the left in Figure 8.1 shows a scatter plot of the average global temperature, in Celsius degrees, for the years 1960-2000. There is no formula for that relationship, but the points seem to trend upwards (on average). So we asked Excel to connect the dots to see

the jagged rise and fall, and then we drew a line on the graph that looked like a reasonable approximation for the trend. The result is on the right. Then we used the line to predict a temperature of 14.58 degrees Celsius for 2010. In fact that average was 14.63 degrees Celsius. Given how up and down the data are (despite the long term average trend) we could hardly expect an accurate prediction. We added the textboxes and the arrows to the spreadsheet to explain how we drew the line.

The line we drew is a *model* — a mathematical construction that approximates something in the real world. This particular model is linear — the line that seems to match the data best. We could have used that model in 2000 to make a prediction for 2010 — an estimate of what the temperature might be in a future year for which we didn't (at the moment) have data.

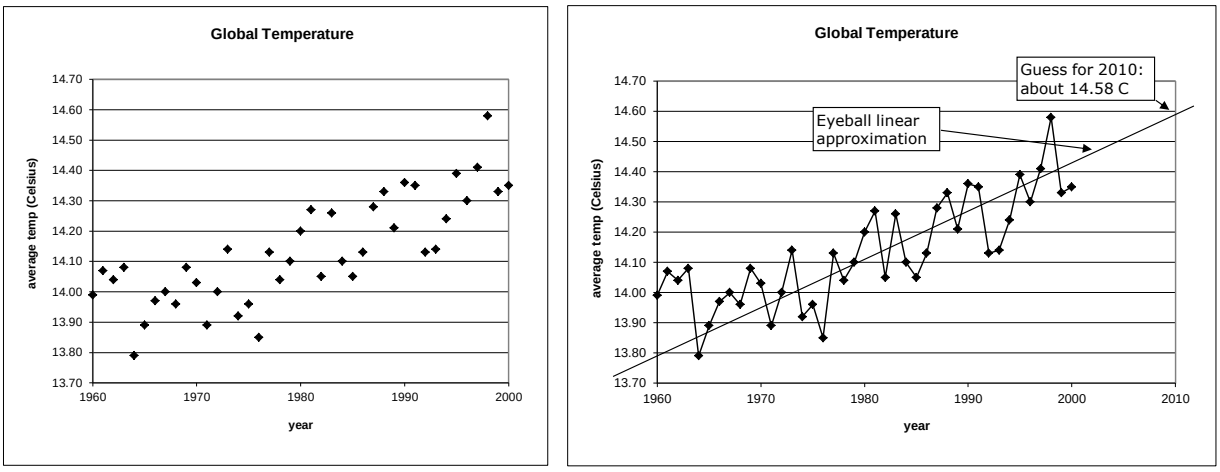


Figure 1.1: Global average temperature, 1960-2000

Excel knows the mathematics for finding the model line we guessed at "by eye". Figure 8.2 shows how invoke it: select the chart, select `Layout` from `Chart Tools`, select `Trendline` and then `Linear Trendline`. Excel draws the second line shown in Figure 8.3.

Not quite. Figure 8.4 shows how to format the trendline: select it (by right clicking); select `Format Trendline ...`; `Forecast Forward` 10 periods (10 years). Check the boxes for `Display Equation` and `Display` R`-squared value` — we will need that data soon. You can change the `Trendline Name`, `Line Color` and `Line Style` if you wish.

Excel calls the line that best fits a scatterplot a trendline. Its official name is *regression line*. We learned (or remembered) in Chapter **??** that straight lines are described by linear equations. The one for the regression line in Figure 8.3 is

$$y = 0.0116x - 8.9214.$$

The slope of the regression line matters most. In this example it says that on average global

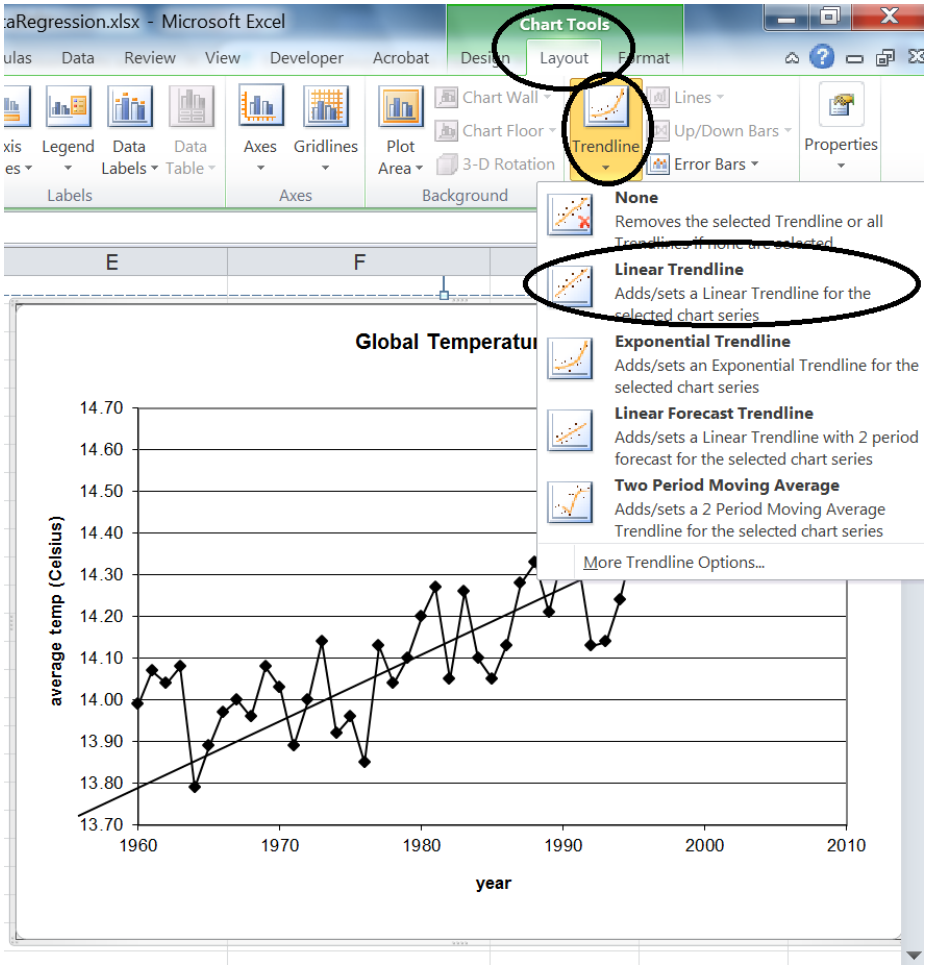


Figure 1.2: Adding a trendline to a chart

temperature is increasing at a rate of

$$0.0116 \ \frac{\text{degrees (Celsius)}}{\text{year}}.$$

That's just over a hundredth of a degree (Celsius) per year, or a tenth of degree per decade. (Remember that the units of the slope are (units of y)/(units of x.)

The intercept for this linear equation, with its units, is

$$-8.9214 \text{ degrees (Celsius)}.$$

Supposedly, that is the temperature predicted (retroactively) by the regression line for year 0. That's nonsense, of course.

In principle, we can use the equation of the line instead of our eyeball approximation to make

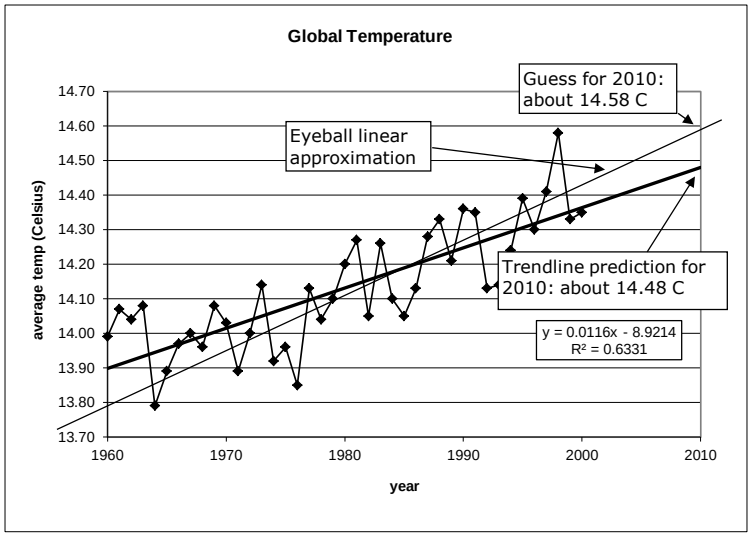


Figure 1.3: Global average temperature, 1960-2000 (prediction to 2010)

our 2010 prediction. If we let $x = 2010$ we find the prediction

$$\begin{aligned} \text{average 2010 temperature} &= y \\ &= 0.0116 \times 2010 - 8.9214 \\ &= 14.2786 \\ &\approx 14.29 \text{ degrees Celsius.} \end{aligned}$$

Something is wrong! When we looked at the trendline that Excel drew, we had an estimate of 14.48 degrees. This calculation is not even close to that estimate. Stop and think: we estimated the 2010 temperature visually, using the Excel trendline, as 14.48 degrees Celsius. When we used the equation for that trendline to calculate the 2010 temperature, we got a number that didn't make sense with what we saw on the graph.

Be skeptical. Always ask whether the numbers from a newspaper or a website or a television commentator — or from a computer program — make sense. This one clearly doesn't. If we dig a little deeper we can see why.

It turns out that Excel rounded off the slope and intercept it showed on the chart. It knows the correct values, but thought all the digits were too ugly to display . To find them, enter the command

```
=SLOPE(
```

in a cell (we used cell `H27`, with a label in `G27`). Excel prompted for

```
=SLOPE(known_y's, known_x's)
```

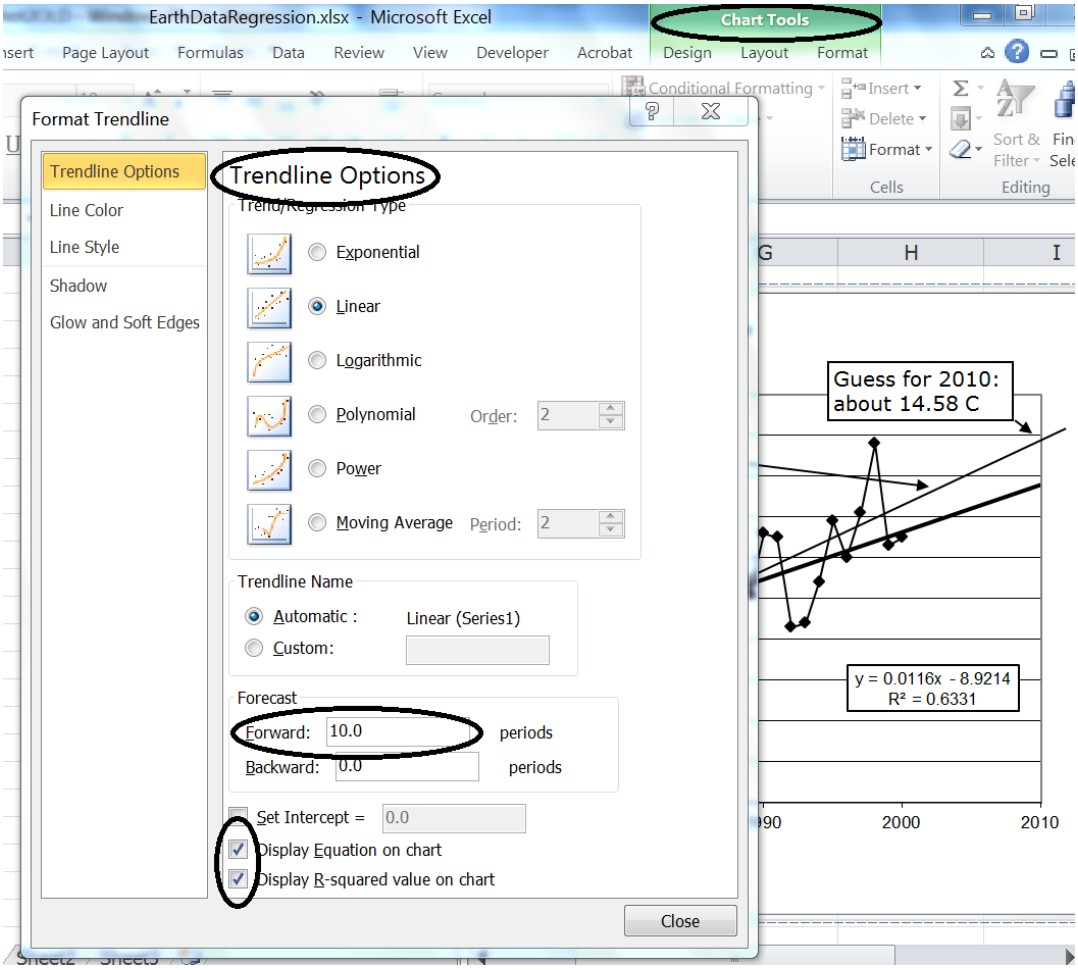



Figure 1.4: Formatting a trendline

so we selected the data

```
=SLOPE(B6:B46,A6:A46)
```

(the years 1960-2000) and Excel told us the correct value: 0.011642857. That's more precise than the rounded value 0.0116 shown on the chart. We found the intercept, -8.921393728, the same way, with the formula `=INTERCEPT(B6:B46,A6:A46)` (in cell `H28`). (In the `SLOPE` and `INTERCEPT` functions the y-values come first and the x-values second, even though in the data table the x-values are first and the y-values second.)

Then the correct equation for the model, before rounding, is

$$y = 0.011642857\ x - 8.921393728.$$

If we set $x = 2010$ in that equation Excel tells us $y = 14.48074913$ (cell `E30`). That rounds

to our visual estimate of 14.48.

We are not the first to discover this problem. Microsoft's support page at `support.microsoft.com/kb/211967` outlines what they call a *workaround*, to show all the decimal places in the trendline equation displayed on the chart. Right click the formula for the trendline on the chart, then select `Format Trendline Label ...`. In the `Number` selection there you can ask for the maximum number of decimal places: 30. Decide for yourself whether you like this method better than asking Excel directly for the `SLOPE` and `INTERCEPT`.

We've said repeatedly that it was wrong to show lots of decimal places when reporting approximate numbers, even when those decimal places appeared in your calculator or spreadsheet. But in this example we saw that too much rounding is wrong too. Using a slope rounded to four significant digits may give a ridiculous answer. The short answer to the question "when should you round?" is

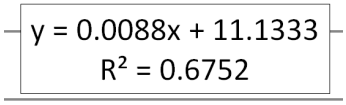
> While you compute, use all the digits you have, even if it's more than you need. Round only when you're done.

Keep this in mind from now on — both when doing the problems in this text and when working with Excel (or another software program) in the future.

1.2 The greenhouse effect

Most climate scientists are convinced that the reason the Earth is warming is the increase in the concentration of greenhouse gases like carbon dioxide in the air.

A greenhouse is warm in the winter because sunlight enters through the glass roof, which prevents the inside air it heats up from escaping. Carbon dioxide (CO_2) behaves similarly in the atmosphere — it lets sunlight in but doesn't let heat out. The chart on the left in Figure 8.5 displays the data and the regression line showing how average temperature varies with the amount of CO_2 in the atmosphere. When e asked Excel to show the equation of the trendline this box appeared on the chart:



.

The slope of the trendline is

$$0.0088 \ \frac{\text{degrees Celsius}}{\text{part per million of } CO_2}.$$

An increase of one part per million of CO_2 corresponds to somewhat under one hundredth of a degree (Celsius) increase in temperature.

That's the trendline slope with four significant digits. If we ask Excel for more we see

$$0.00881808540214804 \ \frac{\text{degrees Celsius}}{\text{part per million of } CO_2}.$$

We would use that value in any computations we made.

The chart on the right in Figure 8.5 shows the increase in CO_2 concentration over the years (it does not mention temperatures at all). There the slope of the regression line is

$$1.3569 \ \frac{\text{parts per million of } CO_2}{\text{year}};$$

on average, every year the CO_2 concentration increases by about 1.36 parts per million.

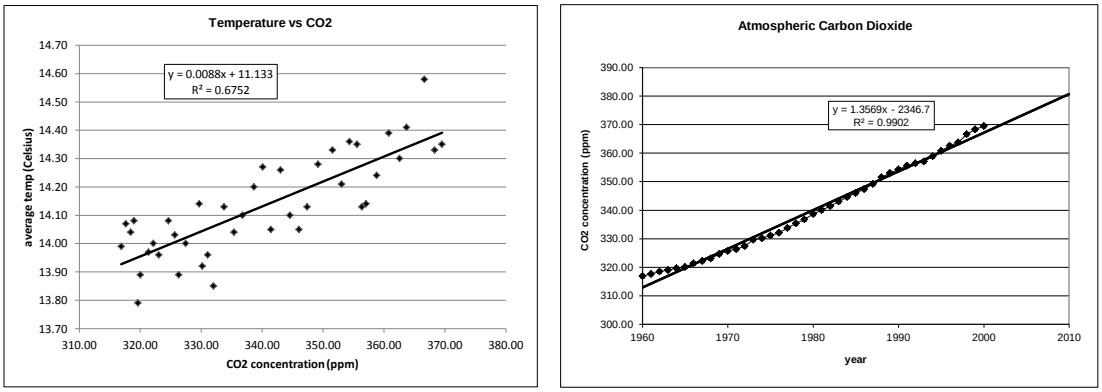


Figure 1.5: CO_2, time and temperature, 1960-2000

The original data set contains three columns of information, listing the year, average global temperature and CO_2 concentration. In Section 8.1 we looked at the relationship between year and average global temperature, which documents the trend called "global warming" in the news. In this section we looked at the other two relationships, between temperature and CO_2 concentration and between CO_2 concentration and time, in hopes of understanding what might be behind the observed temperature trend. In the next two sections we'll think about what we may learn this way.

1.3 How good is the linear model?

How much a regression line helps understand the data and make predictions depends in part on how close the data points are to the line. Common sense tells you that the relationship

between carbon dioxide concentration and time (on the right in Figure 8.5) is likely to be more reliable than that between carbon dioxide and temperature (on the left), which in turn looks better than that between temperature and time (Figure 8.3).

The official statistical measure of "close to the line" is a number between zero and one called "R-squared". The closer R-squared is to 1 the better the regression line fits the data. In Figure 8.3 R^2 is just 0.63321 — not very good. That matches what we can see in the chart — the temperature seems to be increasing on the average, but can go up and down unpredictably from year to year. In the chart on the right in Figure 8.5 the R^2 value is 0.9902, which is very close to 1. In fact the measured 2010 concentration was 389.78 parts per million, so the relative error in the prediction is about -2.5%.

We are being deliberately vague about how close the R^2 should be to 1 to declare that the fit is "good." There are no rules for this. In the exercises below you will have a chance to develop your intuition.

We were careful to use the word "corresponds" when discussing the increase in CO_2 concentration and the increase in average temperature, not the word "causes". The data only say that the CO_2 concentration and the temperature are *correlated* — they trend together. They don't say one causes the other. Data can never tell you that. Climate scientists who work at understanding the physics and chemistry of carbon dioxide in the atmosphere have created scientific models that suggest causation. We will return to this distinction in Section 8.4.

There is much more to the climate change debate: some who accept the scientific models that say that greenhouse gases cause global average temperatures to increase are not convinced that the increase in greenhouse gases is due to human activity, and therefore see no need to change the way we use energy.

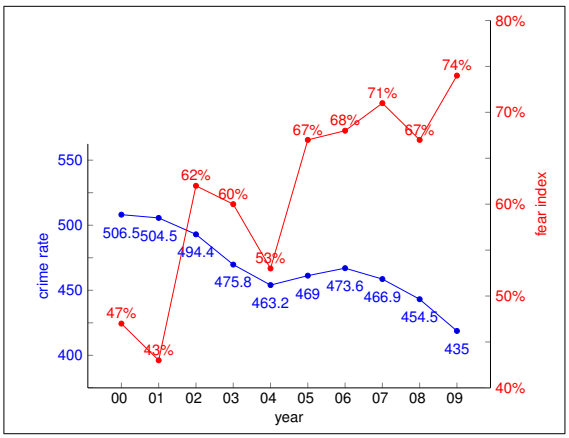
1.4 Regression nonsense

The graphic in Figure 8.6 resembles one that appeared in *The Boston Globe* on January 14, 2010 in a story headlined "Imaginary fiends", which began

> In 2009, crime went down. In fact it's been going down for a decade. But more and more Americans believe it's getting worse. [1]

The data in the accompanying table are from `www2.fbi.gov/ucr/cius2008/data/table_01.html` and `www.gallup.com/poll/123644/Americans-Perceive-Increased-Crime.aspx`. The FBI measures the crime rate in violent crimes per 100,000 people. The fear index is the percentage of people who say crime is going up.

[1] `www.boston.com/bostonglobe/ideas/articles/2010/02/14/imaginary_fiends/`



year	crime rate	fear index
2,000	506.5	47
2,001	504.5	43
2,002	494.4	62
2,003	475.8	60
2,004	463.2	53
2,005	469.0	67
2,006	473.6	68
2,007	466.9	71
2,008	454.5	67
2,009	435.0	74

Figure 1.6: Crime down, fear up

The headline seems to announce a juicy story. The graph is drawn to accentuate the apparent contradiction, since the scales on both y axes don't start at 0. We will use these numbers to illustrate the kinds of nonsense arguments you can make with regression lines. There are three variables to play with: the year, the crime rate, and the fear index. We will focus on them two at a time and imagine different kinds of conclusions.

We started by entering the data in Excel, using the table in the online version of *Common Sense Mathematics* to save typing and prevent typing errors. To do that, select and copy the data from the table. Then paste it into Excel. The bad news is that it is then just text, all in one column. The good news is that Excel can separate the columns of data: open the `Data` tab and select `Text to Columns`. Then entering `Next` on all the dialog windows does the job.

Our work is in the spreadsheet `www.cs.umb.edu/~eb/qrbook/./crimeDropsFearsRise.xlsx` .

The first graph in Figure 8.7 shows a scatterplot and trendline for the last two columns in the table. There we asked Excel to construct a graph with crime rate as the independent variable.

Since we chose crime rate as the independent variable it's easy to look at the graph — and the trend line — and conclude that the increase in crime rate is closely related to the decrease in the fear index. The regression line slopes down — high crime rates seem to come along with decreased fear of crime. The R-squared value is 0.60 — perhaps not compellingly high, but we won't let that stop us from thinking about the data. What might the correlation mean? Could an increase in crime (the independent variable on the x-axis) cause people to be less afraid? Here's an attempt at an explanation: Perhaps when crime is rare it's reported spectacularly in the news and people are frightened, while when it's common it gets less press

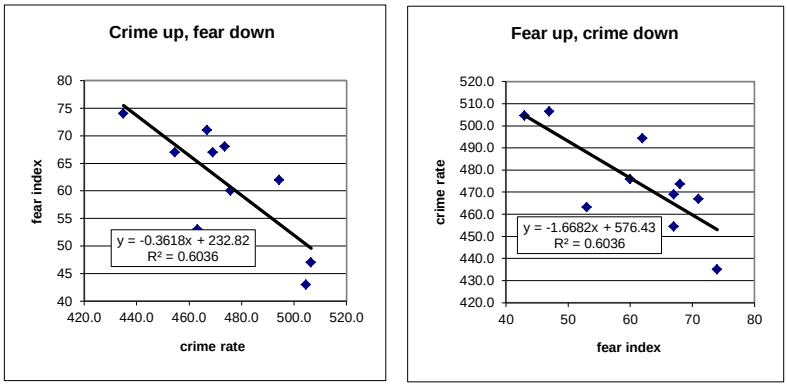


Figure 1.7: Crime vs fear regressions

and most people don't notice it as much because it isn't happening to them.

Does that make sense? Not to us, but it's the kind of argument you frequently see or hear — a simpleminded attempt to explain what seems to be a real "this is true because of that" connection, or perhaps what a politician would like you to believe is a real connection.

The second graph in Figure 8.7 shows the same data with the fear index as the independent variable. That changes our view of the data. Now we see crime dropping as fear increases. How might we explain that? Perhaps we'd argue that increasing fear of crime leads to more pressure on the police to arrest criminals, thus reducing the amount of crime. That's more plausible than the other way around, but still a shallow unconvincing analysis of complex social phenomena. Both the crime rate and the fear of crime are changing over time, one decreasing while the other increases, but just because we can find a trendline doesn't mean either change causes the other.

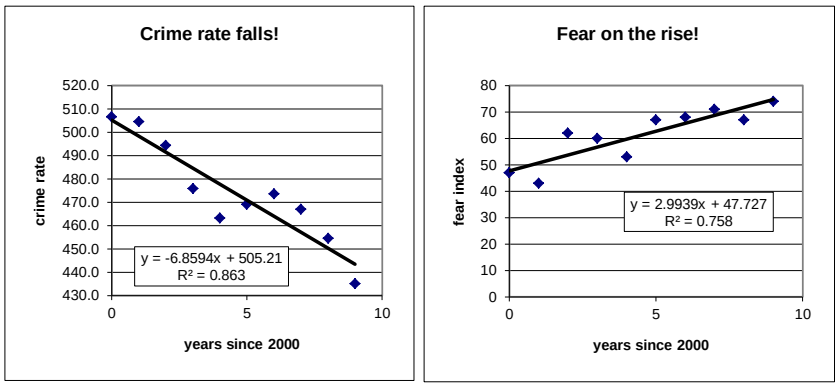


Figure 1.8: Fear index and crime rate over time

We can see the two trends separately if we plot each with time as the independent variable, as in Figure 8.8. With these charts we can create other nonsense arguments. The slope of the fear index regression line is about 3 percentage points per year. Since the index was at 74%

in 2009, if the trend continues then in about 8 more years, in 2017, 98% of the population will believe that crime is getting worse every year. The second regression line says the crime rate is actually falling each year by about 7 violent crimes per 100,000 people, so in 2017 when everyone believes things are getting worse it will be down from 435 to about 380. Neither of these predictions carries much conviction.

The news story that prompted this discussion is misleading in another way. When we found the data on which it is based we discovered that in the previous decade, from 1990 to 2000, the crime rate and the fear index were both decreasing. The author of the article chose not to tell us that. He *cherry-picked* the data to make his point (whatever it is) more dramatic. You can find all the numbers in our spreadsheet at `www.cs.umb.edu/~eb/qrbook/./crimeDropsFearsRise.xlsx` .

The moral of this story:

Correlation is not causation.

It's very easy to use regression to link variables (crime rate and fear index, as in this example), to suggest trends and to make predictions or interpret correlation as explanation. Just because you can doesn't mean you should. It's often wrong. Watch out for people who do.

1.5 Exercises

Exercise 1.5.1. [U][W][Section 8.1][Goal 8.1] A trendline for linear data.

(a) What values would you expect to see for the slope, intercept and R-squared if you were to add a trendline to the Tamworth electricity bill in the spreadsheet `www.cs.umb.edu/~eb/qrbook//ElectricityBill/TamworthElectric.xlsx`?

(b) What would the trendline look like on the graph in Figure **??**?

(c) Add the trendline and verify your predictions.

Exercise 1.5.2. [U][Section 8.3][Goal 8.1] Anscombe's quartet.

> Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. [2]

[2] `en.wikipedia.org/wiki/Anscombe's_quartet`

Use the data in `www.cs.umb.edu/~eb/qrbook/./AnscombesQuartet.xlsx` for the tasks that follow.

(a) For each data set, use Excel to find the mean of the x and y values. Label them in your spreadsheet.

(b) Do the mean values describe these four data sets very well? Explain.

(c) Graph each set of (x, y) values. Label each graph ("data set 1" etc.). Write a sentence or two describing the relationship between the x and y values, using what you see on the graph. Talk about how strong that relationship is (but don't calculate the R-squared value yet).

(d) Display the trendline, the trend line equation and the R^2 value on each graph.

(e) Round the slope and intercept to two decimal places. Write a sentence comparing the slope, intercept and R-squared value for each of the data sets.

(f) Explain in your own words how these examples demonstrate the importance of graphing data before analyzing it.

(g) The short description at the beginning of this problem also talked about the effect of "outliers" on statistical properties. In this context, an outlier is a number that lies outside most of the numbers in the data set. Does each of the data sets contain an outlier? If so, how does that outlier influence the basic statistics for each data set?

(a) For each data set, use Excel to find the mean of the x and y values. Label them in your spreadsheet.

In each case, the mean of the x-values is 9 and the mean of the y-values is 7.50.

(b) Do the mean values describe these four data sets very well? Explain.

Even without graphing these data sets, I can see that while they have the same means, they are somewhat different. The first three data sets share the same x-values, so I'm not surprised that the means of the x-values are the same, but they have different sets of y-values. The fourth data set is the one that is most different from the others. The x-values are almost all equal to 8, which is very different from the first three data sets.

(c) Graph each set of (x, y) values. Label each graph ("data set 1" etc.). Write a sentence or two describing the relationship between the x and y values, using what you see on the graph. Talk about how strong that relationship is (but don't calculate the R-squared value yet).

Data set 1: These points follow a roughly linear relationship and I would expect the correlation to be fairly strong. The points are a bit scattered, but the general trend is that as x increases, so does y.

Data set 2: These points follow more of a curved relationship (I would guess a quadratic relationship) and don't follow a linear path very well. As x increases, y increases up to a point, then decreases. Not sure how the correlation would be.

Data set 3: These points follow a strong linear relationship except for the point (13,12.74) which doesn't follow the pattern. If I ignore that point, I think the correlation would be perfect (that is, the points lie on one line). Even with that point included, I would expect a reasonably strong correlation. I can see that as x increases, y increases for the most part.

Data set 4: This one is just strange. Almost all the points lie on the vertical line through (8, 0) except one point, (19, 12.5), which is far away. I'm not sure what to say here.

(d) Display the trendline, the trend line equation and the R^2 value on each graph.

See below for answers.

(e) Round the slope and intercept to two decimal places. Write a sentence comparing the slope, intercept and R-squared value for each of the data sets.

For each set of points, I found that the trendline equation is $y = 0.50x + 3.0$. And for each data set, the R-squared value is $R^2 = 0.67$. The comparison is easy: the equations and the R-squared values match for each data set.

(f) Explain in your own words how these examples demonstrate the importance of graphing data before analyzing it.

If I had just crunched the numbers, I would not have seen that these are very different sets of data. The graphs are important and add additional information. Drawing a picture (or having Excel draw a picture) is worth the time.

(g) The short description at the beginning of this problem also talked about the effect of "outliers" on statistical properties. In this context, an outlier is a number that lies outside most of the numbers in the data set. Does each of the data sets contain an outlier? If so, how does that outlier influence the basic statistics for each data set?

To understand an outlier, ask yourself if the graph would look very different if you removed a particular point. For data sets 1 and 2, I don't see any obvious point that stands out. The points generally follow the trends described above. Data set 3, however, is different. The point (13, 12.74) lies outside the trend and if we removed it, the remaining points would lie on a line. The point (13, 12.4) is an outlier. For data set 4: if we remove the point (19, 12.5) then the remaining points lie on the vertical line through (8, 0). The point (19, 12.5) is an outlier.

Exercise 1.5.3. [S][Section 8.1][Section 8.2][Goal 8.1] Faster than a speeding bullet.

The spreadsheet at `www.cs.umb.edu/~eb/qrbook/./MarathonWinningTimes.xlsx` shows the history of the winning time in the Boston Marathon for men and women from 1966 (when women first ran) through 2013.

(a) Graph the men's and women's winning times depending on the year, properly label the axes and add a trendline for each data column.

(b) What is the average rate at which the men's finishing time changed from year to year?

(c) Use the trendline to predict when the men's winner will finish in two hours. How confident are you in that prediction?

(d) Use the trendline to predict when the men's winner will finish in one hour. How confident are you in that prediction?

(e) The trendlines suggest that in about six years the fastest woman will be as fast as the fastest man, and will be faster thereafter. Explain why the lines say that, and why it's nonsense.

(f) Make a better prediction about the long run relation between men's and women's winner finishing times.

[See the back of the book for a hint.] Look at the data starting in about 1980.

(a) Graph the men's and women's winning times depending on the year, properly label the axes and add a trendline for each data column.

See `www.cs.umb.edu/~eb/qrbook/./../Answers/MarathonWinningTimes.xlsx` .

(b) What is the average rate at which the men's finishing time changed from year to year?

The slope of the men's trendline is -0.16 minutes per year, or about 10 seconds per year.

(c) Use the trendline to predict when the men's winner will finish in two hours. How confident are you in that prediction?

Extending the trendline shows that it crosses the 120 minute line in about 2055, so about 40 years from now.

The current record is 2:03 (123 minutes). I think the top runner will crack two hours a lot sooner than 40 years from now. If the record drops at an average rate of 10 seconds per year we'll see 120 minutes in only 18-20 years.

(d) Use the trendline to predict when the men's winner will finish in one hour. How confident are you in that prediction?

Another hour off the time would take 360 years at the rate of one minute every six years. That's nonsense. No one will ever run that fast.

(e) The trendlines suggest that in about six years the fastest woman will be as fast as the fastest man, and will be faster thereafter. Explain why the lines say that, and why it's nonsense.

The trendlines cross at about 2018 (four years, not six), at a time of about 127 minutes, or 2:07. I don't believe it. The women's trendline drops unrealistically fast, because of the really steep drops in times when women first started running the marathon.

(f) Make a better prediction about the long run relation between men's and women's winner finishing times.

Hint: look at the data starting in about 1980.

I used Excel to find the women's times relative to the men's, by dividing. The graph of the values starting in about 1981 (yellow cells) show that the women's fastest times are steadily about 10%-20% larger than the men's. I suspect that will continue to be the case, as both records drop.

Exercise 1.5.4. [U][Section 8.1][Goal 8.1] The leaning tower of Pisa.

The famous "Leaning Tower of Pisa" began to lean even while it was under construction in the 1170s. The table in Figure 8.9 shows the measured lean for the years 1975 through 1987.[3]

Year	Lean (m)
1975	2.9642
1976	2.9644
1977	2.9656
1978	2.9667
1979	2.9673
1980	2.9688
1981	2.9696
1982	2.9698
1983	2.9713
1984	2.9717
1985	2.9725
1986	2.9742
1987	2.9757

Figure 1.9: The Tower of Pisa

[3]This picture is from `www.raphaelk.co.uk/web%2520pics/Italy/second/pisa-lina-1.jpg`. The data are from `filebox.vt.edu/users/jemarsh2/LectureNotes/Ch10Examples.pdf`. The second column displays the lean as the distance in meters between where a particular point on the tower would be if the tower were straight and where it actually is.

(a) Construct the regression line for this data and estimate (visually) what the lean was in the year 2000.

(b) How good is that estimate likely to be?

(c) What is the slope of the regression line? What are its units? What does it mean?

(d) Check your estimate using the equation of the regression line. Can you use the formula as it appears in the chart, or do you need more decimal places?

(e) Explain why the actual numbers in the data table for the Tower of Pisa depend on the height of the "particular point" at which measurements were taken. What would the numbers be if the point were twice as high? Would the linear regression line be just as good?

(f) What has happened to the Tower of Pisa since 1987?

Exercise 1.5.5. [S][Section 8.1][Goal 8.1] Beverage consumption.

The spreadsheet at `www.cs.umb.edu/~eb/qrbook/./BeverageConsumption.xlsx` contains data on the amounts of milk, bottled water and soft drinks consumed in the United States between 1980 and 2004.

(a) Use Excel to create a scatter plot of this data. Label the data series and the axes correctly.

(b) Explore correlations among the various categories (for example, between milk and water). Write about what you discover. In particular, which kinds of consumption are most closely correlated?

(c) Use the regression lines to make some predictions for years following 2004.

(d) Find the source of the data in `www.cs.umb.edu/~eb/qrbook/./BeverageConsumption.xlsx` . If you find data for other years there, discuss the validity of your predictions.

[See the back of the book for a hint.] Try a Google search for

Per capita consumption of selected beverages in gallons

.

(a) Use Excel to create a scatter plot of this data. Label the data series and the axes correctly.

The spreadsheet can be found at `www.cs.umb.edu/~eb/qrbook/./BeverageConsumptionSolution.xlsx` .

(b) Explore correlations among the various categories. Write about what you discover. In particular, which kinds of consumption are most closely correlated?

I used the Excel `CORREL()` function to find the correlation coefficients. Then I squared them to find the R-squared values. Here are the results:

Pair	Correlation	R-squared
milk-water	-0.989	0.978
milk-soda	-0.922	0.851
water-soda	-0.884	0.782

That tells me milk and bottled water are most closely correlated. The minus sign means that as the consumption of milk declines the consumption of bottled water increases.

(c) Use the regression lines to make some predictions for years following 2004.

I asked Excel to project the regression lines out to 2010. I then estimated values for 2007 by looking at the graph. (I could have asked Excel to work with the linear function defining the regression line, but decided that the numbers were so inexact that I would just estimate by eye.)

I entered the values in the table in (d) below.

(d) Find the source of the data.

I followed the hint and Googled

Per capita consumption of selected beverages in gallons .

.

The first hit was a link to a spreadsheet at `www.census.gov/compendia/statab/2010/tables/10s0210.xls` that gave figures through 2007, saved locally as `www.cs.umb.edu/~eb/qrbook/./BeverageConsumptionThrough2007.xlsx` .

The following table contains the values for 2007, along with my predictions from the regression lines.

Beverage	2007 prediction	2007 actual
milk	21	20.7
water	24	29.1
soda	58	48.8

The regression line predictions are pretty good for milk and bottled water, but too high for soda. When I look at the data that's not too surprising. Soda consumption seems to have peaked in about 2000 and was level for the next four years. The regression line grows then because it's taking into account the rapid growth between 1980 and 2000. I bet a regression that started with just the 2000-2004 data would predict a value much closer to the 50 gallons that was observed.

Exercise 1.5.6. [S][Section 8.1][Goal 8.1] Energy consumption.

The Excel spreadsheet `www.cs.umb.edu/~eb/qrbook/./EnergyConsumption.xlsx` contains a table showing the annual United States energy consumption, measured in terawatt-hours, between 1949 and 2005.

(a) Insert a new column labeled "years since 1949" in between the years column and the consumption column. Use Excel to fill in the cells for this column.

(b) Use Excel to find a linear trendline for this data. Include the equation and R^2-value for the trendline on the graph.

(c) Is this trendline a good fit for the data?

(d) What is the slope of this line? Include the units in your answer. Use your answer for the slope to complete the sentence: "For every additional year that passes, total energy consumption ..."

(e) Estimate total energy consumption in the years from 2006 to the present.

(f) Look for data with which to check the estimates from the previous part of the exercise.

(a) See the solution spreadsheet at `www.cs.umb.edu/~eb/qrbook/./EnergyConsumptionSolution`

(b) See the spreadsheet.

(c) The trendline is a good fit for the data since $R^2 = 0.9594$, which is very close to 1.

(d) The slope of the trendline is 361.9 twh/year.

For every additional year that passes, total energy consumption increases by about 360 twh.

(e) For 2009 the prediction is about 32,000 twh. It's wrong to report more significant digits than that.

(f) I haven't time to find a good source for the actual 2009 value (yet). Perhaps a student will provide one.

I did discover that U.S. energy consumption actually declined in 2008 and 2009 because of the economic crisis.

Exercise 1.5.7. [S][Section 8.1][Goal 8.1] Supply and demand for office space.

The data in Table 8.10 appeared on page B5 in *The Boston Globe* on April 3, 2010.

quarter	vacancy rate	rent ($/ft^2)
Q1 '06	11.8%	38.76
Q1 '07	7.5%	47.54
Q1 '08	6.0%	62.20
Q1 '09	9.0%	49.24
Q1 '10	11.1%	42.46

Table 1.10: Less in rent, more in vacancy

(a) Build and then discuss a linear regression line for the dependence of rent per square foot on vacancy rate.

(b) How do your conclusions change when you adjust rents to take inflation into account?

Figure **??** shows the charts for both parts of the exercise.

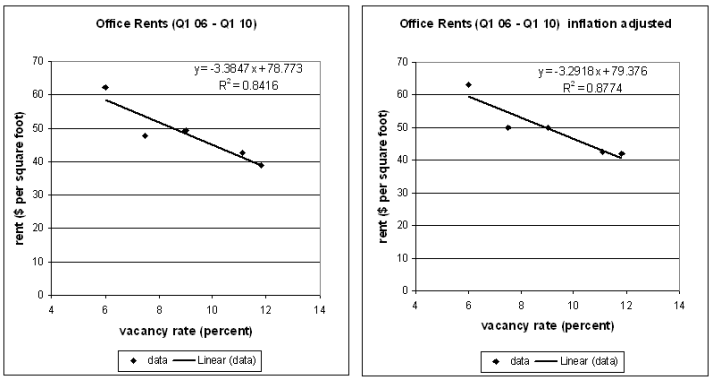


Figure 1.11: Office rents, Q1 06 - Q1 10

(a) Build and then discuss a linear regression line for the dependence of rent per square foot on vacancy rate.

The regression line has a slope of -3.38. That means that each 1% increase in the vacancy rate corresponds to a decrease of $3.38 per square foot in office space rent.

R^2 is 0.84, which means the correlation is pretty good.

(b) How do your conclusions change when you adjust rents to take inflation into account?

After I used the Bureau of Labor Statistics inflation calculator to write all the rents in 2010 dollars, the slope was -3.29 and the R^2 was 0.88. That's a little higher.

Exercise 1.5.8. [S][Section 8.1][Goal 8.1] [Goal 8.3] Office rents.

On February 22, 2008 *The Boston Globe* ran a story under the headline "Office rents reach dizzying heights" that featured graphs like those in Figure 8.13.

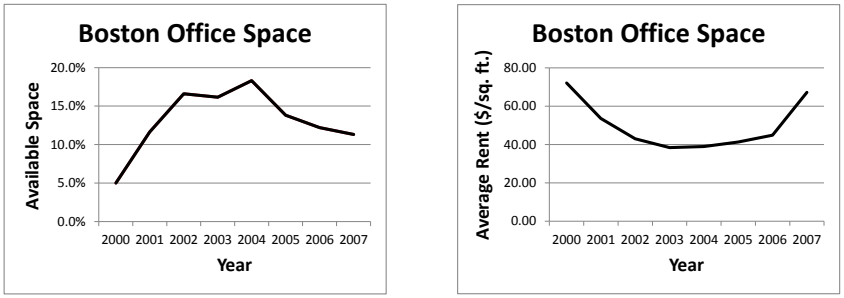


Figure 1.12: Boston office rental rates

The shapes of the curves illustrate the law of supply and demand — the more space is available the less you have to pay for it.

You can find the data in the spreadsheet `www.cs.umb.edu/~eb/qrbook/./BostonOfficeRents.xlsx` .

(a) Show how rental cost depends on the percent of space available by creating a scatter plot using columns `D` and `F` and a regression line for that scatter plot. Identify the slope and its units. How good is the correlation?

(b) Use the graph and the formula to estimate office rent when the availability rate is 8%.

(c) The spreadsheet contains data on the vacancy rate as well as the availability rate. Create a scatterplot illustrating how the vacancy rate depends on the availability rate. Add a regression line and discuss what it tells you.

See `www.cs.umb.edu/~eb/qrbook/./../Answers/BostonOfficeRentsSolution.xlsx`

(a) Show how rental cost depends on the percent of space available by creating a scatter plot using columns `D` and `F` and a regression line for that scatter plot. Identify the slope and its units. How good is the correlation?

The slope of the line (with units) is

$$-2.76 \frac{\text{\$/square foot}}{\text{percentage point of vacancy rate}}$$

It tells me that for each increase of one percentage point in the availability rate the average rent falls by about $2.76 per square foot.

R^2 is about 0.76, which is OK but not wonderful.

(b) Use the graph and the formula to estimate office rent when the availability rate is 8%.

The picture suggests that the rent will then be about $64 per square foot. The formula says

$$-2.7572 \times 8 + 86.096 = 64.0384$$

which rounds to 64. My guess from the graph was pretty good!

(c) The spreadsheet contains data on the vacancy rate as well as the availability rate. Create a scatterplot illustrating how the vacancy rate depends on the availability rate. Add a regression line and discuss what it tells you.

The slope of the regression line is

$$0.7 \frac{\text{percentage point of vacancy}}{\text{percentage point of availability}}$$

The picture shows that is not a really good fit over the whole range. At higher availability rates (12% to 18%) the vacancy rate seems to be pretty constant at 9%.

Exercise 1.5.9. [U][Section 8.3][Goal 8.1] [Goal 8.1] First class mail.

Table 8.14 shows the cost of sending first class mail weighing up to one ounce.

(a) Copy and paste the data into Excel, then draw a graph of the data.

(b) Insert the trendline and display the trendline equation and the R-squared value on the graph.

(c) Write a sentence interpreting the slope of the trendline.

(d) Is this a strong correlation? Explain.

Exercise 1.5.10. [U][Section 8.1] Speed vs. MPG, revisited.

Exercise **??** looked at the relationship between speed and fuel consumption. You can do this problem even if you didn't do that one.

(a) Read data from the graph in Figure **??** and enter it in Excel.

Year	Cost (cents)
1976	13
1978	15
1981	18
1985	22
1988	25
1991	29
1995	32
1999	33
2001	34
2002	37
2006	39
2007	41
2008	42
2009	44
2012	45
2013	46

Table 1.13: First class mail

(b) The information cited in Exercise **??** states that for each 5 mph you drive over 50 mph, your decrease in fuel economy means that you pay an additional \$0.25 for gas. Use Excel to graph the data corresponding to speeds above 50 mph. Construct a regression line for this data. What does the slope of the regression line tell you about how fuel economy changes as speed increases? If your speed increases by 5 mph, how does your fuel economy change, on average?

(c) Use Excel to convert the data in your table from mpg to gallons per 100 miles. Graph the data again and insert the regression line. What does the slope of the regression line tell you about how fuel economy changes as speed increases? Is it easier to explain how fuel economy changes when your speed increases by 5 mpg?

Exercise 1.5.11. [S][Section 8.1][Goal 8.1] College costs.

The spreadsheet `www.cs.umb.edu/~eb/qrbook/./CollegeCosts2010.xlsx` shows the annual mean cost for tuition and fees at private and public four-year colleges in the U.S. between 1999 and 2010.

(a) Create a properly labeled graph showing how mean private and public education costs changed in the years 1999-2010.

Insert a linear trendline for each set of data. Use Excel to forecast the trendline out to 2015 (that is, 16 years past 1999).

(b) Write the equation for private education costs.

(c) Write the equation for public education costs.

(d) Interpret the numerical value of the slope in each trendline equation. That is, write a sentence explaining what the slope represents.

(e) Use your trendline equations to determine the projected mean tuition cost at both private and public four year colleges for 2015.

(f) Compare your answers from the previous questions with the graph. Are the answers consistent or do you need to use more digits in your calculation?

(g) In Chapter **??**, Exercise **??** presents data on public and private college spending increases. Compare the data there with the revenue increases from tuition and fee data here.

(a) Build chart with trendlines.

See `www.cs.umb.edu/~eb/qrbook/./../Answers/CollegeCosts2010Solution.xlsx` .

(b) Write the equation for private education costs.

Excel says:

```
y = 1102.7x + 14918
```

(c) Write the equation for public education costs.

```
y = 397.3x + 3072.6
```

(d) Interpret the numerical value of the slope in each trendline equation.

The slopes of the trendlines show that the cost of public education is increasing at a rate of $397 per year while that for private education is increasing at a rate of $1103 per year.

(e) Use your trendline equations to determine the projected mean tuition cost at both private and public four year colleges for 2015.

2015 is 16 years from 1999, so I plugged 16 into each of the equations and project that then public college education will cost $32,561 while private will cost just $21,275.

The R-squared value for each of the trendlines is very close to 1, so I am pretty confident about these predictions.

(f) Compare your answers from the previous questions with the graph. Are the answers consistent or do you need to use more digits in your calculation?

The answers match up well enough.

(g) In Chapter **??**, Exercise **??** presents data on public and private college spending increases. Compare the data there with the revenue increases from tuition and fee data here.

Exercise 1.5.12. [U][Section 8.1][Goal 8.1] Manhattan rental market.

On February 22, 2008 *The Boston Globe* ran a story under the headline "Office rents reach dizzying heights" that featured graphs like those in Figure 8.13.

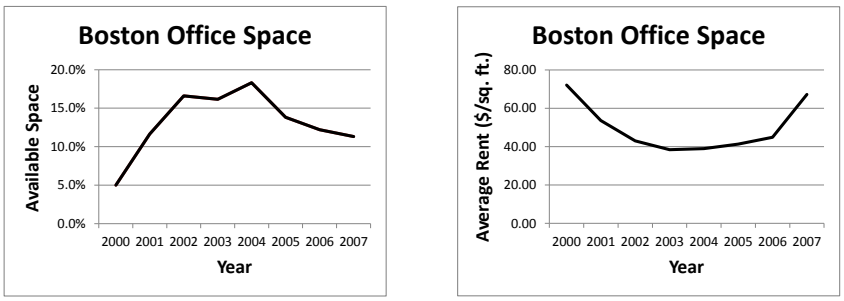


Figure 1.14: Boston office rental rates

The shapes of the curves illustrate the law of supply and demand — the more space is available the less you have to pay for it.

You can find the data in the spreadsheet `www.cs.umb.edu/~eb/qrbook/./BostonOfficeRents.xlsx` .

(a) Show how rental cost depends on the percent of space available by creating a scatter plot using columns `D` and `F` and a regression line for that scatter plot. Identify the slope and its units. How good is the correlation?

(b) Use the graph and the formula to estimate office rent when the availability rate is 8%.

(c) The spreadsheet contains data on the vacancy rate as well as the availability rate. Create a scatterplot illustrating how the vacancy rate depends on the availability rate. Add a regression line and discuss what it tells you.

See `www.cs.umb.edu/~eb/qrbook/./../Answers/BostonOfficeRentsSolution.xlsx`

(a) Show how rental cost depends on the percent of space available by creating a scatter plot using columns `D` and `F` and a regression line for that scatter plot. Identify the slope and its units. How good is the correlation?

The slope of the line (with units) is

$$-2.76 \frac{\$/\text{square foot}}{\text{percentage point of vacancy rate}}.$$

It tells me that for each increase of one percentage point in the availability rate the average rent falls by about $2.76 per square foot.

R^2 is about 0.76, which is OK but not wonderful.

(b) Use the graph and the formula to estimate office rent when the availability rate is 8%.

The picture suggests that the rent will then be about $64 per square foot. The formula says

$$-2.7572 \times 8 + 86.096 = 64.0384,$$

which rounds to 64. My guess from the graph was pretty good!

(c) The spreadsheet contains data on the vacancy rate as well as the availability rate. Create a scatterplot illustrating how the vacancy rate depends on the availability rate. Add a regression line and discuss what it tells you.

The slope of the regression line is

$$0.7 \frac{\text{percentage point of vacancy}}{\text{percentage point of availability}}$$

The picture shows that is not a really good fit over the whole range. At higher availability rates (12% to 18%) the vacancy rate seems to be pretty constant at 9%.

Exercise 1.5.13. [U][Section 8.3][Goal 8.1] [Goal 8.1] First class mail.

Table 8.14 shows the cost of sending first class mail weighing up to one ounce.

(a) Copy and paste the data into Excel, then draw a graph of the data.

(b) Insert the trendline and display the trendline equation and the R-squared value on the graph.

(c) Write a sentence interpreting the slope of the trendline.

(d) Is this a strong correlation? Explain.

Exercise 1.5.14. [U][Section 8.1] Speed vs. MPG, revisited.

Exercise **??** looked at the relationship between speed and fuel consumption. You can do this problem even if you didn't do that one.

(a) Read data from the graph in Figure **??** and enter it in Excel.

Year	Cost (cents)
1976	13
1978	15
1981	18
1985	22
1988	25
1991	29
1995	32
1999	33
2001	34
2002	37
2006	39
2007	41
2008	42
2009	44
2012	45
2013	46

Table 1.15: First class mail

(b) The information cited in Exercise **??** states that for each 5 mph you drive over 50 mph, your decrease in fuel economy means that you pay an additional $0.25 for gas. Use Excel to graph the data corresponding to speeds above 50 mph. Construct a regression line for this data. What does the slope of the regression line tell you about how fuel economy changes as speed increases? If your speed increases by 5 mph, how does your fuel economy change, on average?

(c) Use Excel to convert the data in your table from mpg to gallons per 100 miles. Graph the data again and insert the regression line. What does the slope of the regression line tell you about how fuel economy changes as speed increases? Is it easier to explain how fuel economy changes when your speed increases by 5 mpg?

Exercise 1.5.15. [S][Section 8.1][Goal 8.1] Manhattan rentals.

Figure 8.15 appeared in *The New York Times* on October 15, 2011. [4] The data are in Table 8.16, from `www.citi-habitats.com/`.

(a) Reproduce the charts in Figure 8.15 in Excel. Label them properly. If you can, play around with the vertical axis in Excel to make your graphs look like the graphs in the figure.

[4] `www.nytimes.com/2011/10/16/realestate/rents-in-manhattan-rebound-to-record-highs.html`

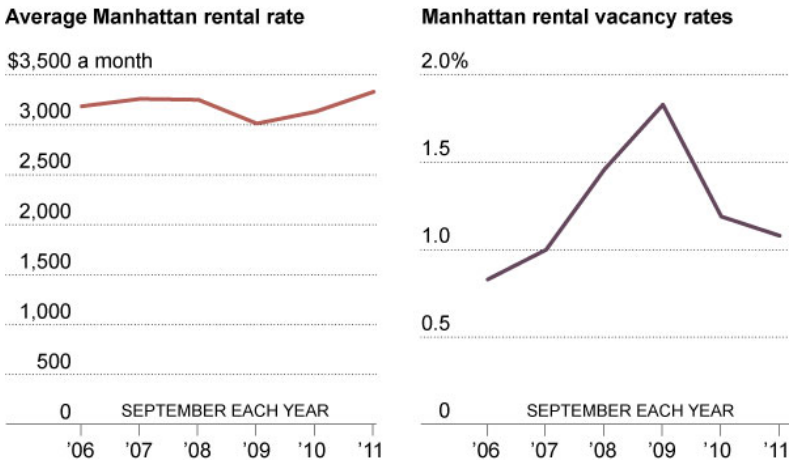


Figure 1.16: Manhattan rental data

year	vacancy rate (%)	monthly rent ($)
06	0.849	3173
07	1.007	3254
08	1.413	3256
09	1.841	3010
10	1.191	3144
11	1.095	3343

Table 1.17: Apartment rents in Manhattan

(b) Create a scatterplot from the second and third columns in Table 8.16, draw a trendline and discuss the correlation between vacancy rate and average monthly rent.

Exercise 1.5.16. [S][Section 8.4][Goal 8.1] Playing with regression lines.

Use the spreadsheet `www.cs.umb.edu/~eb/qrbook/./PlayWithRegression.xlsx` to explore the following questions.

(a) What happens when all the y-values are the same?

(b) What if all but one of the y-values are the same and you vary that one?

(c) What if y decreases as x increases?

(d) What if the x and y values match?

(a) What happens when all the y-values are the same?

I changed the value in cell `B14` to 1 to make all the y-values the same. The trendline equation turned into

$$y = 1.$$

That makes sense, since the slope is 0 and the y-intercept is 1. Excel complains about R^2 and refuses to calculate it.

(b) What if all but one of the y-values are the same and you vary that one?

I changed the value in `B14` from 3 to 4, then to 100.

The single high point kept pulling up the trendline, so its slope got bigger (and its intercept got smaller).

The R^2 value didn't change.

(c) What if y decreases as x increases?

For $x = 1, 2, 3, 4, 5$ I used the values $y = 10, 8, 4, 6, 2$. The trendline had slope -2, which did not surprise me. The correlation was -1; the minus sign was telling me that the line sloped down. Since all the points lie on the line, the values are perfectly correlated and $R^2 = -1$.

(d) What if the x and y values match?

I let $x = 1, 2, 3, 4, 5$ and let $y = 1, 2, 3, 4, 5$. Notice that all the points lie on the trendline. Excel calculates $R^2 = 1$, which makes sense since the line matches up exactly with the points.

Exercise 1.5.17. [U][Goal 8.1] [Section 8.4] Should businesses use private jets?

On May 26, 2012 *The Boston Globe* published a letter to the editor from David V. Dinneen, Executive director of the Massachusetts Airport Management Association. It said in part

> According to a recent report, annual earnings of S&P companies that use general aviation were 434 percent higher than those that did not. [5]

"... use general aviation" is Dineen's way of saying that they have their own fleet of corporate jets.

Explain how and why he is using the statistic he quotes to encourage readers to confuse correlation with causation.

Exercise 1.5.18. [S][Section 8.4][Goal 8.3] Cherry-picking.

[5] `www.bostonglobe.com/opinion/letters/2012/05/25/despite-its-many-benefits-corporate-use-aircraft-sti`
`mbQ6mINMQXbAayzWvFn6NI/story.html`

In Section 8.4, we discovered that the author had "cherry-picked" the data. Find out what "cherry-picking" means, and where the phrase comes from. Find and discuss some examples.

From Wikipedia (reliable in this case)

> Cherry picking, suppressing evidence, or the fallacy of incomplete evidence is the act of pointing to individual cases or data that seem to confirm a particular position, while ignoring a significant portion of related cases or data that may contradict that position.

Exercise 1.5.19. [S][Section 8.4][Goal 8.1] [Goal 8.3] Watch TV! Live Longer!

The data in the spreadsheet `www.cs.umb.edu/~eb/qrbook/./TVData.xlsx` show the life expectancy in years for several countries, along with the number of people per television set in those countries. (The idea (and the data) for this problem come from the article `www.amstat.org/publications/jse/v2n2/datasets.rossman.html`.)

(a) Which countries have the highest and lowest life expectancy at birth? Which have the highest and lowest number of people per television set?

(b) Use Excel to create a properly labelled scatter plot of the life expectancy and people per television data. Find the trendline and display the equation and the R-squared value on your graph.

(c) What is the slope of the trendline (with its units)? Explain its meaning in a sentence.

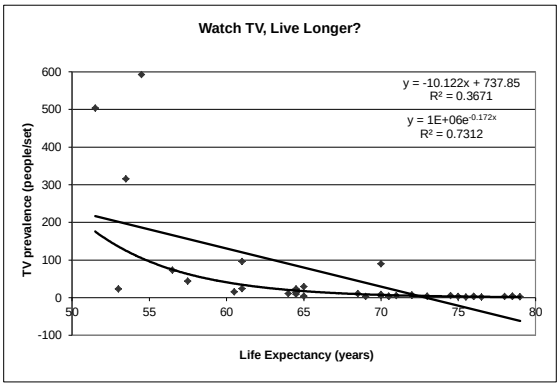
(d) Does a small number of people per television set improve health? Would people in countries with low life expectancy live longer if we sent them shiploads of television sets?

(e) Does living longer increase the number of television sets? If we improved the life expectancy in a country by providing better medical care would that cause there to be fewer people per television set?

(f) What else could be going on here? Why might high life expectancy be strongly correlated with a low ratio of people per tv set?

(a) Life expectancy varies from 79 years in Japan to 51.5 years in Ethiopia. Television prevalence varies from 1.3 people per set in the United States to 592 per set in Myanmar. If you try to find the largest and smallest values by simply scanning the columns of figures you're likely to make a mistake. It's best to sort in Excel.

(b) Here's a correct solution, with an exponential trendline as well as the linear one.



(c) The slope of the trendline is -10.122 (people per TV) per (year of life expectancy). It seems to say that for each decrease of 10 people per TV, life expectancy increases by one year. The correlation isn't very good. The R-squared value is just 0.3671.

The last three questions are all concerned with the same issue. What might account for the fact that longer life expectancy seems to go along with more television sets? The simple answer is that each trend is a consequence of affluence. The richer a society, the better medical care it offers its citizens and the more they have the leisure and the means to watch television.

Exercise 1.5.20. [S] [W] Crime rates revisited.

(a) Use the data in `www.cs.umb.edu/~eb/qrbook/./crimeDropsFearsRise.xlsx` to redo the analysis for the entire period from 1990 to 2009.

(b) Are the crime rates in this exercise consistent with those in the example we studied in Chapter **??**?

[See the back of the book for a hint.] For the second question, all you can really look for is the order of magnitude. If that doesn't match, try to explain why.

(a) Use the data in `www.cs.umb.edu/~eb/qrbook/./crimeDropsFearsRise.xlsx` to redo the analysis for the entire period from 1990 to 2009.

I copied all four charts and changed the data series in each to use the numbers in rows 50:66. (I changed the titles and the scales on the axes too.) Figure **??** shows the result. Now there is a positive correlation. Fear and crime more or less rise and fall together. But the data are scattered and the correlation is quite weak: R^2 is just 0.45. Plotting each variable over time, you can see the crime rate falling quite consistently following a linear trend ($R^2 \approx 0.91$) while fear goes down and then up (the regression line is useless). That might have made an even more interesting news story.

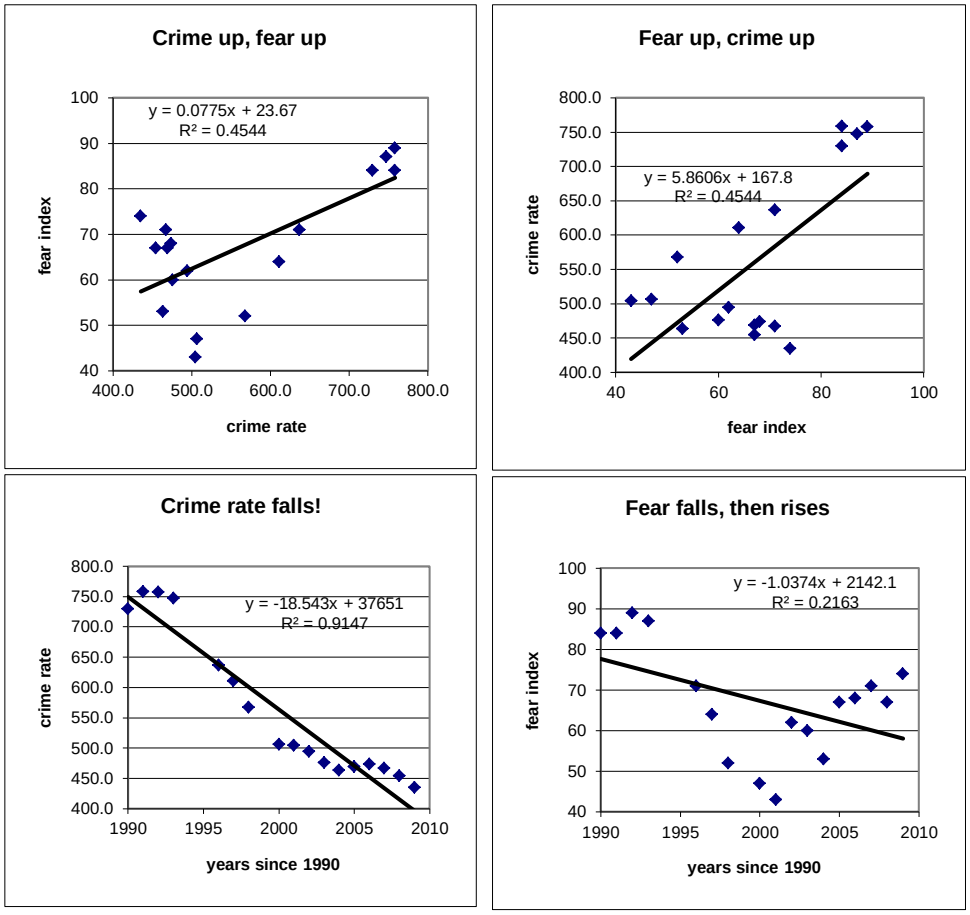


Figure 1.18: Crime:fear correlation

(b) Are the crime rates in this exercise consistent with those in the example we studied in the chapter on units?

The crime rates here are on the order of 500 per 100,000 people. In the discussion in the units chapter they are on the order of 10 per 1000 people. That converts to 1000 per 100,000 people, which is twice as much. Perhaps that's because the ones here are "violent crimes" while the ones there are just "crimes".

Exercise 1.5.21. [U][Section 8.4] [Goal 8.1] [Goal 8.3] The Mississippi River.

> In the space of one hundred and seventy-six years the Lower Mississippi has shortened itself two hundred and forty-two miles. That is an average of a trifle over one mile and a third per year. Therefore, any calm person, who is not blind or idiotic, can see that in the Old Oolitic Silurian Period, just a million years ago next November, the Lower Mississippi River was upwards of one million three hundred thousand miles long, and stuck out over the Gulf of

> Mexico like a fishing-rod. And by the same token any person can see that seven hundred and forty-two years from now the Lower Mississippi will be only a mile and three-quarters long, and Cairo and New Orleans will have joined their streets together, and be plodding comfortably along under a single mayor and a mutual board of aldermen. There is something fascinating about science. One gets such wholesale returns of conjecture out of such a trifling investment of fact.
>
> Mark Twain
> Life on the Mississippi
> `www.gutenberg.org/files/245/245.txt`

Discuss this linear model for the length of the Mississippi river. What's the slope? Can you verify Twain's arithmetic?

Since $242/176 = 1.375$, Twain is right to say the rate is "an average of a trifle over one mile and a third per year." That's the slope.

Projecting that trend backwards, a million years ago the Mississippi would have been about one and one third million miles longer than it was when Twain wrote about it. The Gulf of Mexico is only about 560 miles from north to south (`www.epa.gov/gmpo/about/facts.html`) so the river would have done much more than stick out over the Gulf like a fishing rod — it would have reached more than four times the distance to the moon.

Projecting forward 742 years, the Mississippi would be about 1000 miles shorter. From Cairo (Illinois) to New Orleans is only about 600 miles, so that estimate doesn't seem right. Cairo and New Orleans would be together in only $600/1.375 \approx 400$ years.

Exercise 1.5.22. [U][Section 8.4][Goal 8.3] Well, maybe.

Explain the joke in the cartoon in Figure 8.17 from `xkcd.com/`.

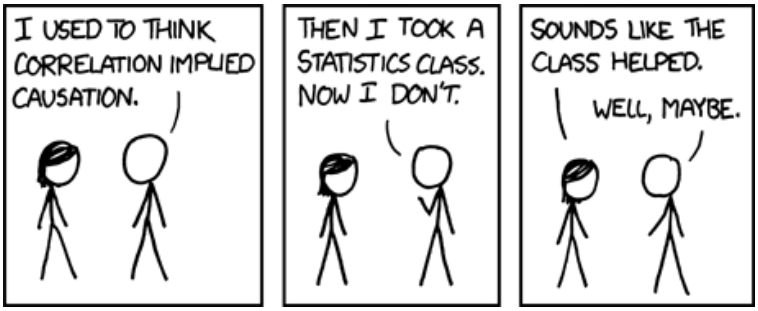


Figure 1.19: Well, maybe.

Exercise 1.5.23. [U][Section 8.1][Goal 8.1] Visiting the White House.

On April 15 2012 *The New York Times* published Figure **??**

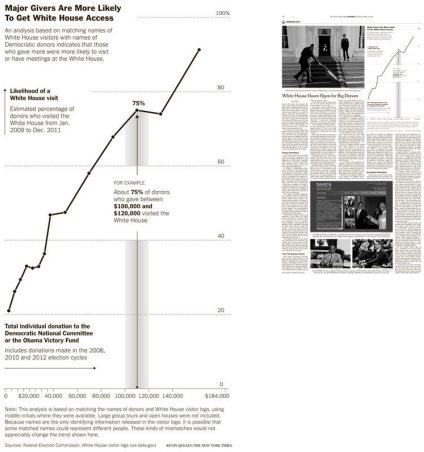


Figure 1.20: Odds of an invitation to the White House

Fit a linear trendline to this data to predict the size of donation that would guarantee an invitation to visit the White House.

You can do this with a ruler and get a good-enough approximate answer. No need to put the data into Excel.

Exercise 1.5.24. [N] It's too darn hot.

In his blog in 2012 Andrew Gelman posted on the topic

2% per degree Celsius . . . the magic number for how worker productivity responds to warm/hot temperatures (`andrewgelman.com/2012/09/persistently-reduced-labor-productivity-may-be-one-`

Exercise 1.5.25. [U][C][Section 8.1][Goal 8.1] Polarization.

Note: The data need work before we can ask the students to deal with them.

Figure **??** appeared in *The Boston Globe* on November 6, 2010. [6] We extracted the numerical data from the graph; you can find it at `www.cs.umb.edu/~eb/qrbook/./polarization.csv` .

(a) Find the trendline modeling a linear relationship between the income share of the top 1 percent of the population and the political polarization index.

(b) Find the trendline modeling a linear relationship between the income share of the top 1 percent of the population in a year and the political polarization index four years earlier.

[6] `www.boston.com/news/nation/washington/articles/2010/11/06/election_opens_up_a_gaping_divide/`

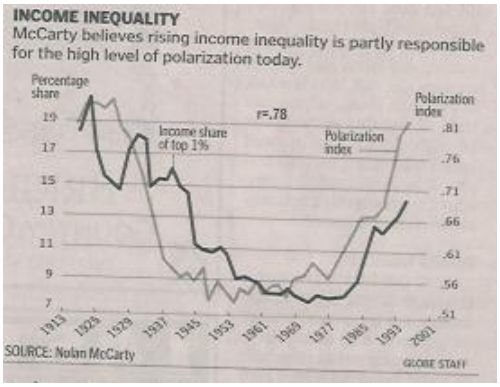


Figure 1.21: Income disparity and political polarization

Exercise 1.5.26. [U][C] Do the math on overrides.

Barry Bluestone and Anna Gartsman wrote an op-ed with that headline in *The Boston Globe* on June 4, 2010. Implicit in what they propose are several linear dependencies among statistics describing towns in Massachusetts:

> We decided to test this theory by simulating the impact on home values of a change in school spending due to a Prop. 2½ override, controlling for other factors. We obtained data on housing values in 2005, the change in housing values between 2005 and 2010, and two measures of perceived school quality: school-wide SAT scores and per pupil expenditures. We found complete data for 176 of the 351 cities and towns in the Commonwealth.
>
> According to our analysis, which controlled for initial home value in 2005, a municipality with SAT scores and per pupil spending levels 20 percent higher than average experienced a 24 percent increase in nominal home value between 2005 and 2010. In contrast, a municipality with SAT scores and per pupil spending 20 percent below average experienced a loss in home value of 11 percent.
>
> So, how much difference would the passage of the Hull override have potentially meant for home values in that community? Hull's 2005 average home value of $366,343 was near the mean for the communities in our study. The average SAT score in the Hull public schools was 961 compared with an average of 1047 for all the study communities. Average per pupil expenditure in Hull was $11,491, some $1,500 higher than average. Based on our home value model, the predicted increase in home values in Hull between 2005 and 2010 was 3.85 percent.
>
> Now what would likely have happened to the average home value in Hull if the recent proposed $1.9 million override had been passed back in 2005? This tax increase would have cost the average homeowner in Hull $506 per year. Over five years, it would have totaled $2,530. However, that tax increase would have resulted in an additional $1,442 spent per pupil. This increase would result in a predicted increase in home value of 6.57 percent rather than the increase of 3.85

percent. The difference between the two predicted values results in an average increase in home value in Hull of \$9,970. [7]

These figures are probably the result of a regression study. Identify the slopes of the regression lines involved, and verify the predictions.

Exercise 1.5.27. [N] Climate changes. climate change

The Economist published Figure **??** on May 12, 2010, along with the following paragraph:

How global surface temperature, ocean heat and atmospheric CO2 levels have risen since 1960.

The record of atmospheric carbon-dioxide levels started by the late Dave Keeling of the Scripps Institute of Oceanography is one of the most crucial of the data sets dealing with global warming. When the measurements started in 1959 the annual average level was 315 parts per million, and it has gone up every year since. To begin with it went up by roughly one part per million per year. Now it is more like two parts per million per year. The figure for 2011 is 391.6. More carbon dioxide in the atmosphere means a stronger greenhouse effect, and various measurements speak to this. Global surface temperature records show a warming over the same period, though because of fluctuations in the climate, air pollution, volcanic eruptions and other confounding factors the rise is nothing like as smooth. A steadier rise can be seen in the heat content of the oceans, measured in terms of the energy stored, rather than the temperature.

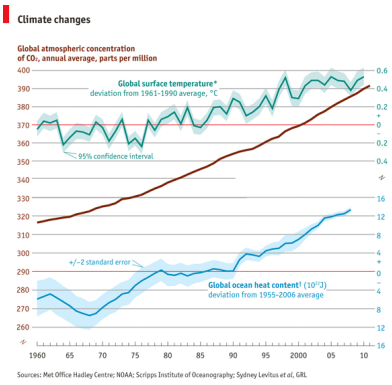


Figure 1.22: Climate change

Exercise 1.5.28. [U][Section 8.2] 2013 Carbon Dioxide.

On September 10, 2014 The Associated Press reported that

[7] www.boston.com/bostonglobe/editorial_opinion/oped/articles/2010/06/24/do_the_math_on_overrides/

> The heat-trapping gas blamed for the largest share of global warming rose to worldwide concentrations of 396 parts per million last year, the biggest year-to-year change in three decades, the World Meteorological Organization said in an annual report.
>
> That's an increase of 2.9 parts per million from the previous year and is 42 percent higher than before the Industrial Age, when levels were about 280 parts per million.
>
> Based on the current rate, the world's carbon dioxide pollution level is expected to cross the 400 parts per million threshold by 2016, said organization Secretary General Michel Jarraud. That is way beyond the 350 amount that some scientists and environmental groups promote as a safe level and that was last seen in 1987. [8]

How might the data in this quotation change the discussion in Section 8.2?

Exercise 1.5.29. [N] Start with a graph.

This is a placeholder for a suggestion from an early reviewer.

> I'd like to see more homework problems here that begin from a graph and trend line, rather than beginning from a data set. Given that it is so easy to mislead with graphs, this would help students to develop those "defensive reading" skills that appear to be one of the goals of this chapter.

Exercise 1.5.30. [U][N] Heart attack risk

At the website `www.cardiosource.org/en/Science-And-Quality/Practice-Guidelines-and-Quality-2013-Prevention-Guideline-Tools.aspxi` you can download a spreadsheet with which to predict your risk of a heart attack. You fill in some values (like your age, blood pressure and cholesterol count) and the spreadsheet tells you your risk.

The formulas it uses are hidden, but you can figure out something about them by experimenting.

For example, try filling in all the fields, then vary the total cholesterol count while keeping all the other values the same. Record the results in another spreadsheet, and produce a graph showing how risk depends on that variable. Is it linear? Approximately linear?

Do the same for some of the other variables.

[8] `hosted.ap.org/dynamic/stories/E/EU_UNITED_NATIONS_GLOBAL_WARMING?SITE=AP&SECTION=HOME&TEMPLATE=DEFAULT&CTIME=2014-09-09-15-29-18`

