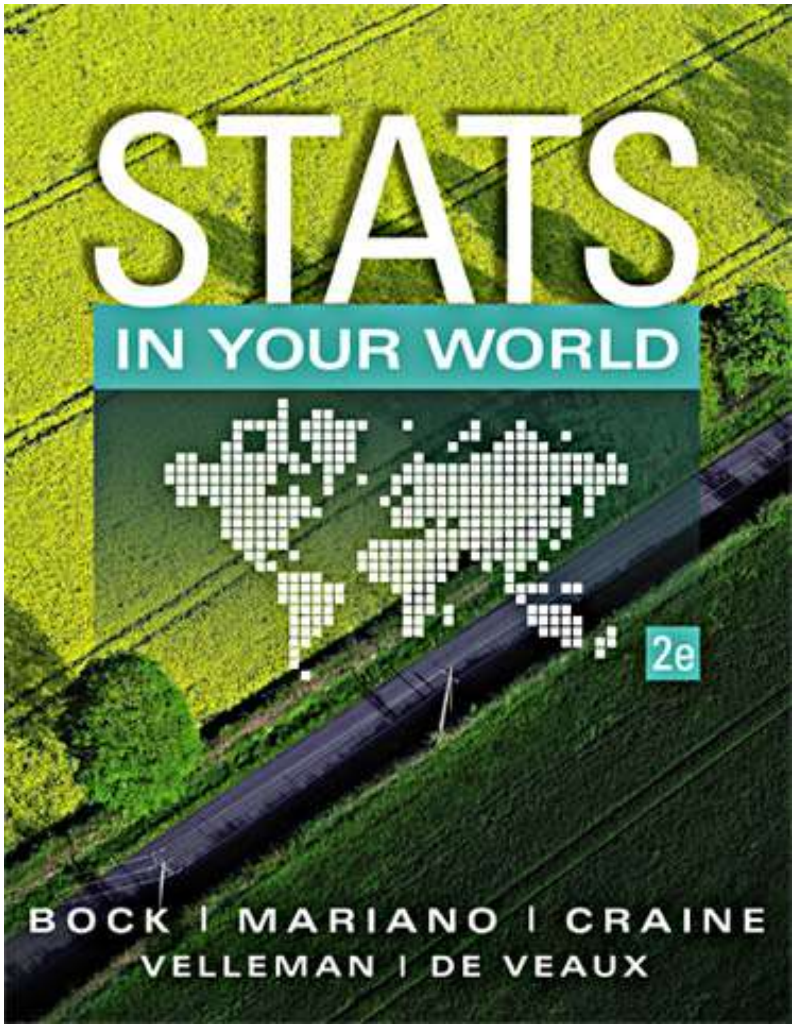
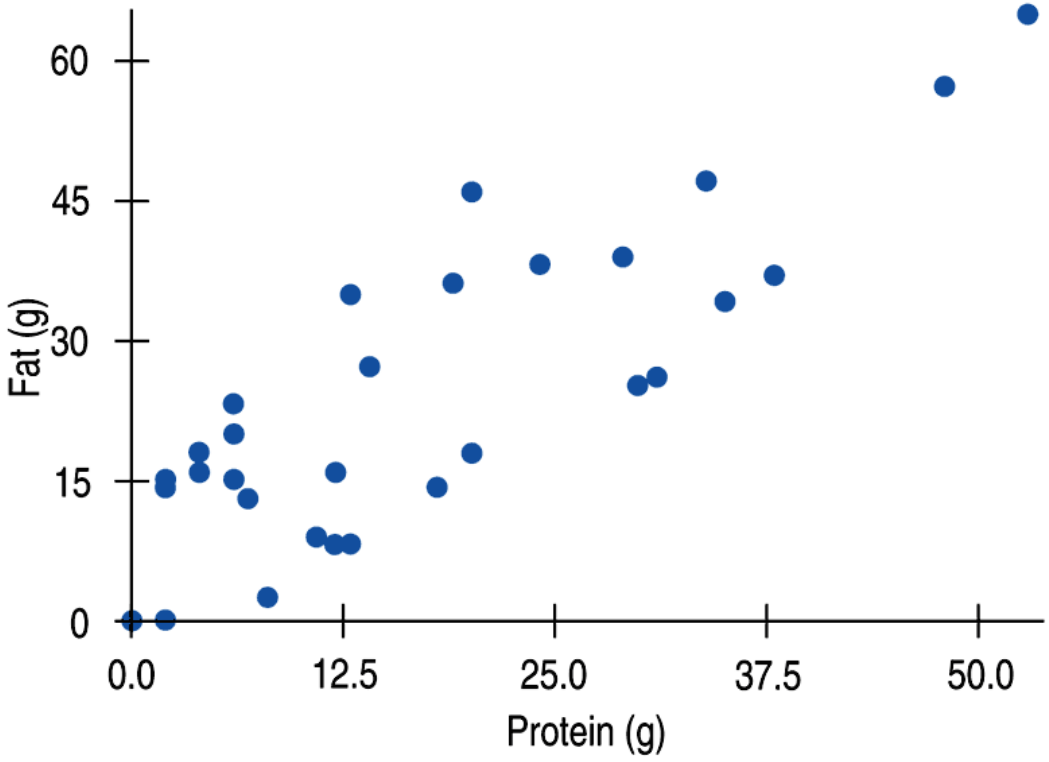
Chapter 7

What's My Line?



Fat Versus Protein: An Example

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



The Linear Model

- The correlation in this example is 0.83. Using that value and the graph, last chapter we said "There is a fairly strong, linear, positive association between protein and fat."

- In this chapter we will say more about the linear relationship between two quantitative variables with a linear model.

The Linear Model (cont.)

- The linear model is just an equation of a straight line through the data.
  - The points in the scatterplot don't all line up, but a straight line can summarize the general pattern with only a couple of parameters.
  - The linear model can help us understand how the values are associated.

Residuals

- The model won't be perfect, regardless of the line we draw.
- As the statistician George Box said, "All models are wrong, but some models are useful."
- Some points will be above the line and some will be below.
- The estimate made from a model is the predicted value, denoted as $\hat{y}$ (y - hat).

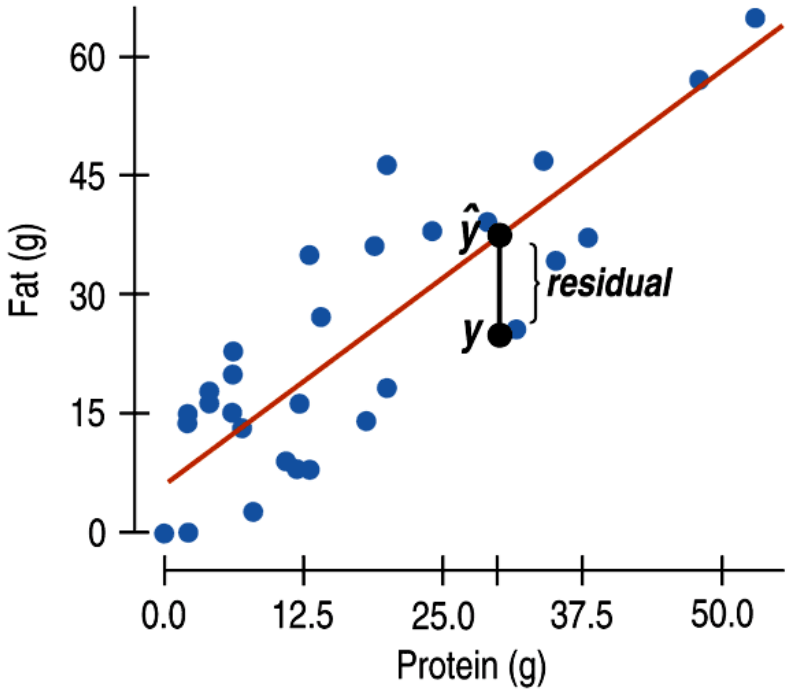
Residuals (cont.)

- The difference between the observed (or true) value and the line's predicted value is called the residual.
- To find the residuals, we always subtract the predicted value from the observed one:

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

Residuals (cont.)

- A negative residual means the predicted value's too big (an overestimate).
- A positive residual means the predicted value's too small (an underestimate).
- In the figure, the estimated fat of the BK Broiler chicken sandwich is 36 g, while the true value of fat is 25 g, so the residual is –11 g of fat.



The Linear Model

- Remember from Algebra that a straight line can be written as:
$$y = mx + b$$
- In Statistics we use a slightly different notation:
$$\hat{y} = a + bx$$
- We write $\hat{y}$ to emphasize that the points that satisfy this equation are just our *predicted* values, not the actual data values.
- This model says that our *predictions* from our model follow a straight line.
- If the model is a good one, the data values will scatter closely around it.

The Linear Model (cont.)

- We write b and a for the slope and intercept of the line.
- b is the slope, which tells us how rapidly $\hat{y}$ changes with respect to x.
- a is the y-intercept, which tells where the line crosses (intercepts) the y-axis.

Best Fit Line

- In our model, we have a slope (b):
  - The slope is built from the correlation and the standard deviations:

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

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

Best Fit Line (cont.)

- In our model, we also have an intercept (a).
  - The intercept is built from the means and the slope:

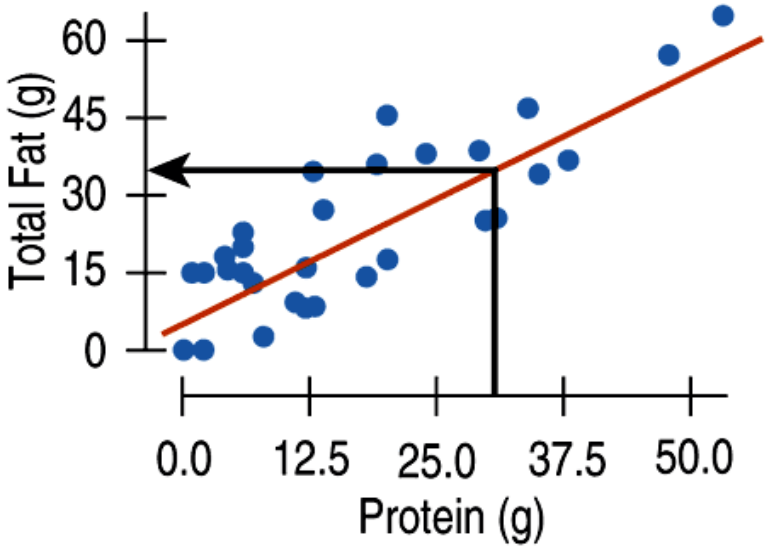
$$a = \bar{y} - b\bar{x}$$

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

Fat Versus Protein: An Example

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

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



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

The Least Squares Line (cont.)

- Since regression and correlation are closely related, we need to check the same conditions for regressions as we did for correlations:
  - Quantitative Variables Condition
  - Straight Enough Condition
  - Outlier Condition

Residuals Revisited

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

$$\textit{Data} = \textit{Model} + \textit{Residual}$$

or (equivalently)

$$\textit{Residual} = \textit{Data} - \textit{Model}$$

Or, in symbols,

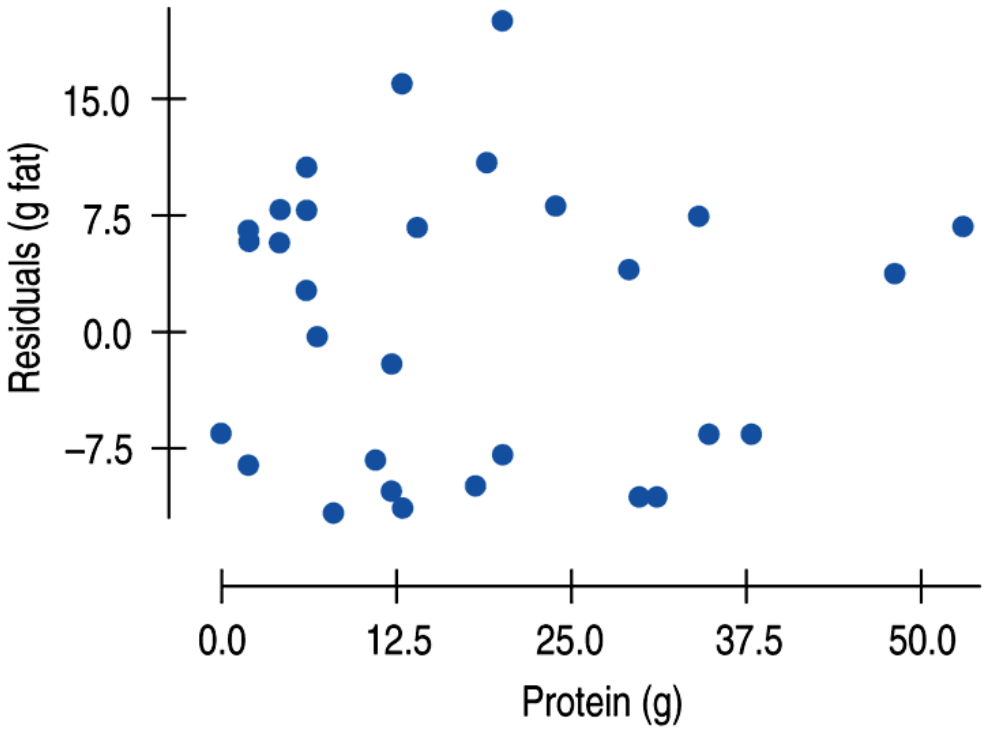
$$e = y - \hat{y}$$

Residuals Revisited (cont.)

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

Residuals Revisited (cont.)

- The residuals for the BK menu regression look appropriately boring:



Assumptions and Conditions

- Quantitative Variables Condition:
  - Regression can only be done on two quantitative variables (and not two categorical variables), so make sure to check this condition.
- Straight Enough Condition:
  - The linear model assumes that the relationship between the variables is linear.
  - A scatterplot will let you check that the assumption is reasonable.

Assumptions and Conditions (cont.)

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

Assumptions and Conditions (cont.)

- Outlier Condition:
  - Watch out for outliers.
  - Outlying points can dramatically change a regression model.
  - Outliers can even change the sign of the slope, misleading us about the underlying relationship between the variables.
- If the data seem to clump or cluster in the scatterplot, that could be a sign of trouble worth looking into further.

A Tale of Two Regressions

- You might be tempted to use a regression equation backwards: plug in a *y*-value and predict an *x*-value.
- But that doesn't work. Our equation is not built to minimize predictions in the *x*-direction.
- If you want to make predictions about the *y*-variable, you need to swap the variables and find a new regression equation.

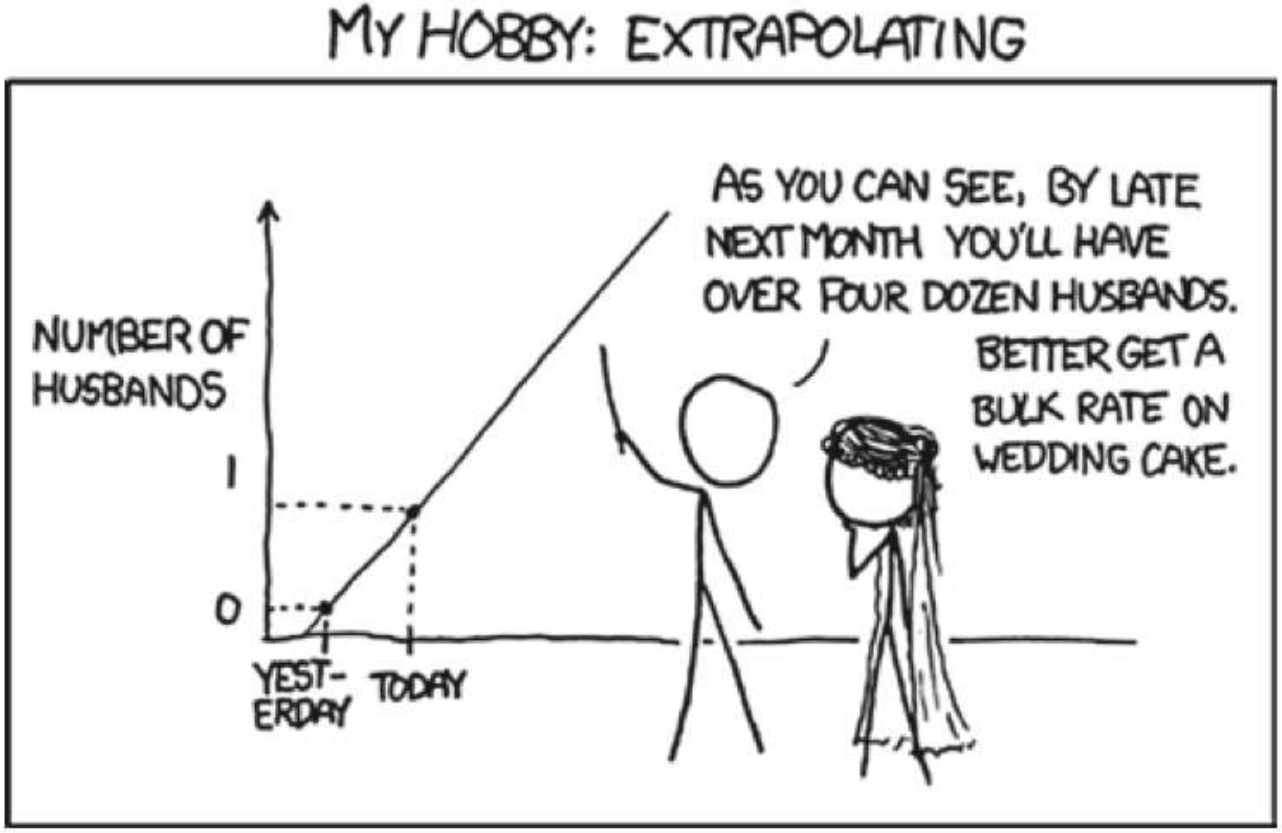
Extrapolation: Reaching Beyond the Data

- Linear models give a predicted value for each case in the data.
- We cannot assume that a linear relationship in the data exists beyond the range of the data.
- The farther the new *x* value is from the mean in *x*, the less trust we should place in the predicted value.
- Once we venture into new *x* territory, such a prediction is called an extrapolation.

Extrapolation (cont.)

- Extrapolations are dubious because they require the additional—and very questionable — assumption that nothing about the relationship between *x* and *y* changes even at extreme values of *x*.
- Extrapolations can get you into deep trouble. You're better off not making extrapolations.

Extreme Extrapolation!

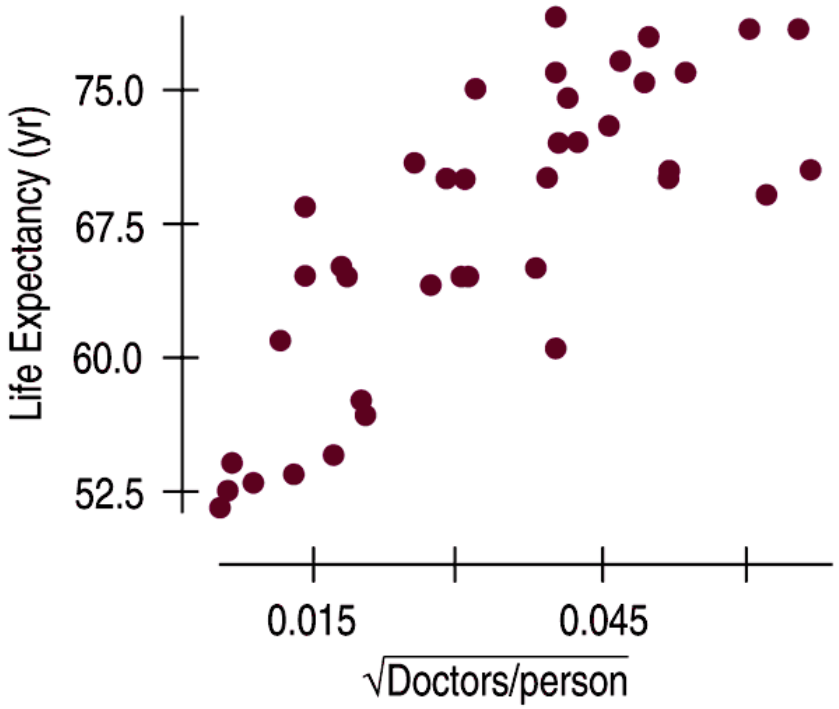


Lurking Variables and Causation

- No matter how strong the association, no matter how straight the line, there is no way to conclude from a regression alone that one variable *causes* the other.
  - There's always the possibility that some third variable is driving both of the variables you have observed.
- With observational data, as opposed to data from a designed experiment, there is no way to be sure that a lurking variable is not the cause of any apparent association.

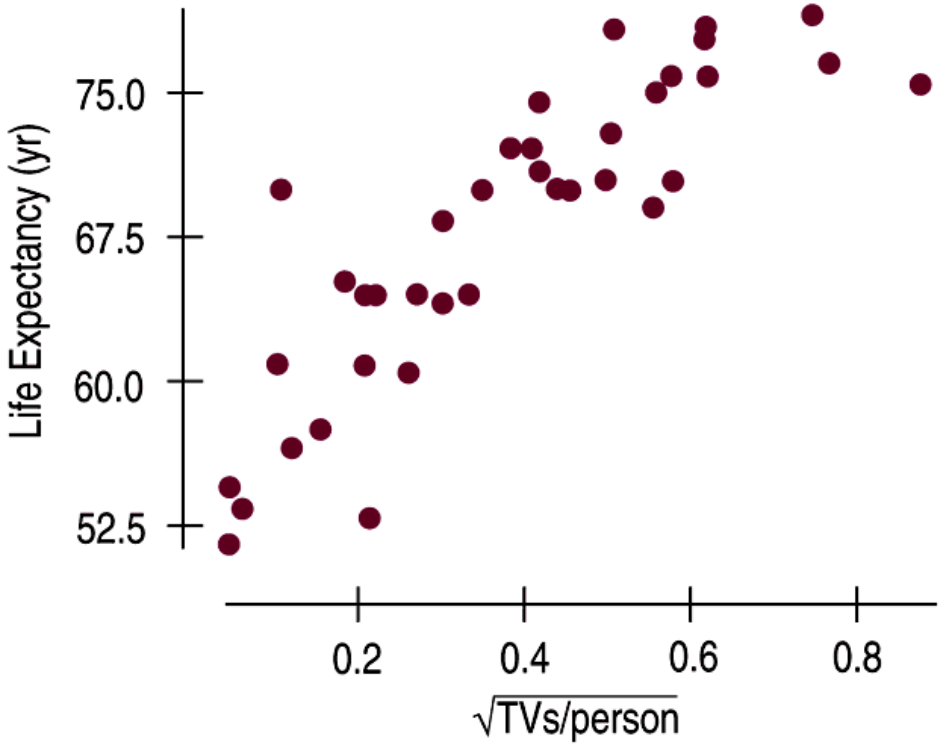
Lurking Variables and Causation (cont.)

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



Lurking Variables and Causation (cont.)

- This new scatterplot shows that the average *life expectancy* for a country is related to the number of *televisions* per person in that country:



Lurking Variables and Causation (cont.)

- Since televisions are cheaper than doctors, send TVs to countries with low life expectancies in order to extend lifetimes. Right?
- How about considering a lurking variable? That makes more sense…
  - Countries with higher standards of living have both longer life expectancies *and* more doctors (and TVs!).
  - If higher living standards *cause* changes in these other variables, improving living standards might be expected to prolong lives and increase the numbers of doctors and TVs.

What Can Go Wrong?

- Make sure the relationship is straight enough.
- Don't fit a straight line to a nonlinear relationship.
- Don't extrapolate beyond the data—the linear model may no longer hold outside of the range of the data.
- Beware especially of extrapolating into the future!
- Beware of lurking variables—and don't assume that association is causation.
- Don't infer that *x* causes *y* just because there is a good linear model for their relationship—association is *not* causation.
- Don't even *imply causation*.