PMU 199H-L0411: "Lies, damned lies, and statistics"

November 11, 2014

Required for November 25

- Homework 2 is due November 25, 10.10 a.m.
- Be prepared to describe briefly your plans for your essay: topic, sources, questions

In the news

- "The rise of Mean World Syndrome" Globe & Mail, Nov. 6.
  "When we slavishly engage in social media to follow big events, the hours spent scrolling have an emotional impact". Although the phrase was invented in the 1970s, well before social media, by "U.S. communications professor George Gerbner, ... Its a cognitive bias wherein consumers of mass media can come to believe that the world is more dangerous than it actually is through constant exposure to violent imagery or commentary". The article also mentions the "emotional contagion" experiment that Facebook ran in June, and was heavily criticized for. The article also mentions a study from UC San Diego, which " found that something as simple as mentions of rainfall on Facebook affected other users moods negatively".
- "How a privacy advocate plans to corrupt online advertisers' data" Globe & Mail, Nov. 6.
  "A new browser extension called AdNauseam will launch next week at the Digital Labor conference at the New School in New York. ... The program works in the background of a Web browser, clicking on ads indiscriminately and boggling the information that ad networks collect."
- "Will MOOCs be flukes" New Yorker, Nov. 7. This article reviews a number of assessments of Massive Open Online Courses, discussing things like completion rate, pass rate, and student satisfaction.

"I hear they're desperate for new users."

Technical Note: Linear Regression (see also §10.4 in *STS*)

In the class activity on diamond prices, you will compute a simple linear regression of price (in dollars) on weight (in carats: 1 carat = 0.2 grams). Each of you has a different set of data, and all the data was randomly generated based on a regression from a real data set. The source of the data is a full page advertisement placed in the Straits Times newspaper issue of February 29, 1992, by a Singapore-based retailer of diamond jewelry, and is available here.

The line you will draw is the least squares line for a plot of price (on the y axis) against quality (on the x axis). The variable chosen for the y axis is usually called the *dependent* variable: it is the variable we would like to predict, or summarize, as a function of the variable that is on the x axis.

The regression line that is drawn has an *intercept* where it crosses the y axis, and a *slope*: remember $y = mx + b$? We will use the notation of *STS*:

$$y = a + bx;$$

the slope is b and the intercept is a. The formulas for the estimates are:

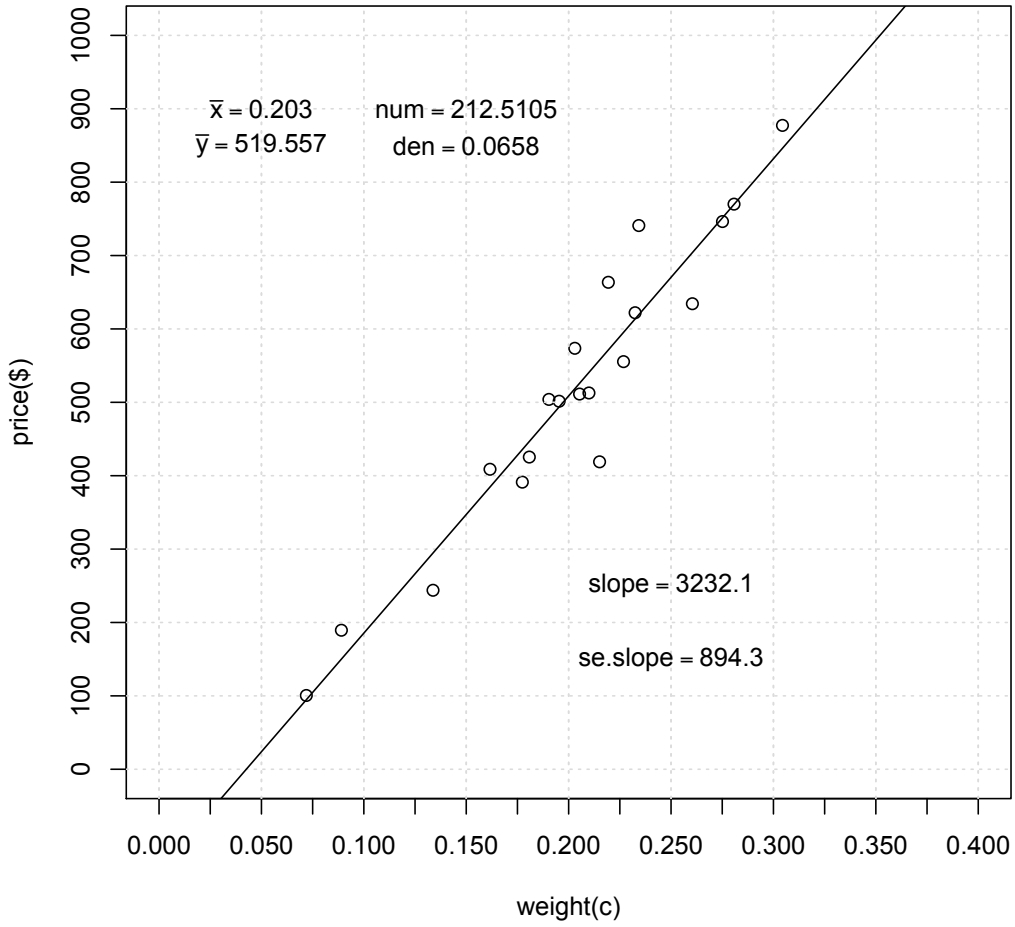
$$\begin{aligned} b &= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\ a &= \bar{y} - b_1\bar{x}, \quad \text{where} \\ \bar{y} &= \sum y_i/n, \quad \bar{x} = \sum x_i/n. \end{aligned}$$

The pairs of points (x_i, y_i) are the (carat,price) data points; there are 20 of them in each data set.

We might expect for example that the price of a diamond could also depend on other features, such as "clarity", or "fire", or perhaps what mine it came from, and so on. *Multiple regression* models the effect of several variables $x_1, \ldots, x_p$ on y, using the formula

$$y = a + b_1x_1 + b_2x_2 + \ldots + b_px_p.$$

Of course, if the data don't follow a line, at least on average, then linear regression doesn't make much sense. However, linear regression is a pretty reliable and simple technique in a lot of cases, and it is a building block for more complicated settings. The study on rainfall and Facebook mentioned on p.1 used such a generalization, 'instrumental variables regression'.



WelcomeHomeSweepstakes Mod ↗ Robin McConnell · a month ago

There are 13,009 prizes and only 315,000 tickets available. 1 chance in 315,000 tickets x 13,009 draws equates to approx. 1 chance in 25 to win a prize.

Reply · Share ›