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

November 25, 2014

Required for December 2

- Your essay is due, in class preferred, but email submission accepted until midnight.
- Attendance is required, s.v.p. I’ll give a survey of the world of “Big Data”.

In the news

- “A persuasive chart showing how persuasive charts are” NY Time Upshot, Nov 24. Blog Post by Justin Wolfers.

  Turns out there was a study published recently in the journal *Public Understanding of Science*, in which subjects were told:

  > “A large pharmaceutical company has recently developed a new drug to boost peoples’ immune function. It reports that trials it conducted demonstrated a drop of 40 percent (from 87 to 47 percent) in occurrence of the common cold. It intends to market the new drug as soon as next winter, following F.D.A. approval.”

  Then, half the subjects were randomly assigned to also see a (very uninformative) graph of exactly the same information.

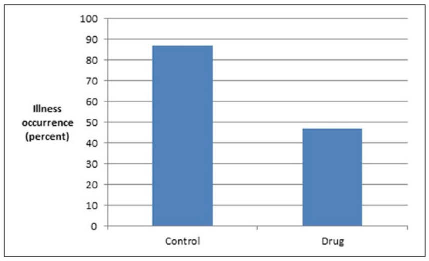
  

  Figure 1. Graph displayed with Study 1

  It turned out that subjects that saw the chart, as well as the summary above, were more likely to think the medication was effective.

  > “a higher percentage believed the medication would truly reduce illness for the graphs group (96.55%) than for the control group (67.74%)”

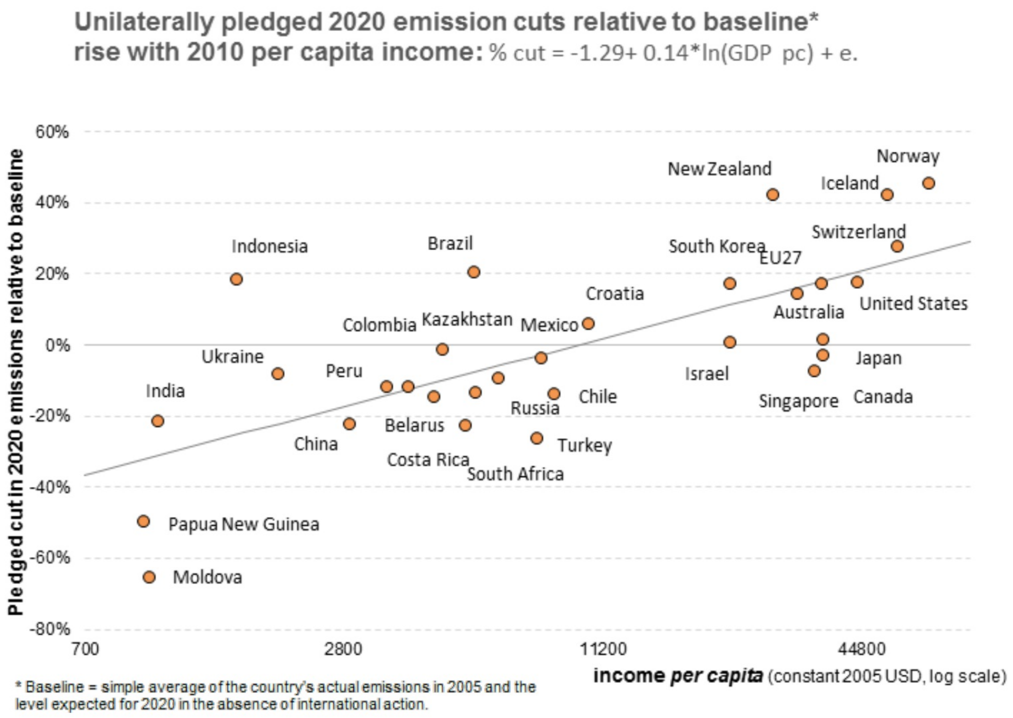
  You can read the original study here.

- A tweet from @TimHarford: “Which countries should curb their carbon emissions most? What’s fair?”, led me to an article on the website VoxEU.org, a policy portal of the Centre for Economic Policy Research.[1]. The article, “Fair

[1] “Vox aims to promote research-based policy analysis and commentary by leading scholars”

shares in pledged carbon cuts", discusses a 'scorecard' approach to see how much carbon cuts each country 'should' offer. They conclude that based on their proposed principles, countries are making cuts in line with expectations.

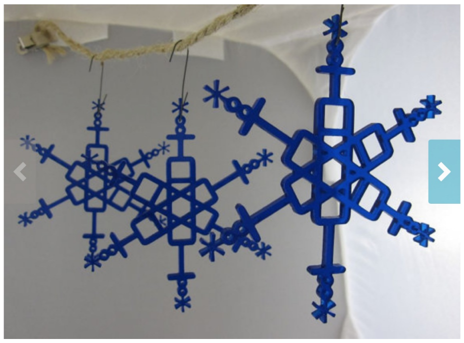
Figure 1. Pledged cuts are again larger for higher-income countries



Unilaterally pledged 2020 emission cuts relative to baseline*
rise with 2010 per capita income: % cut = -1.29+ 0.14*ln(GDP pc) + e.

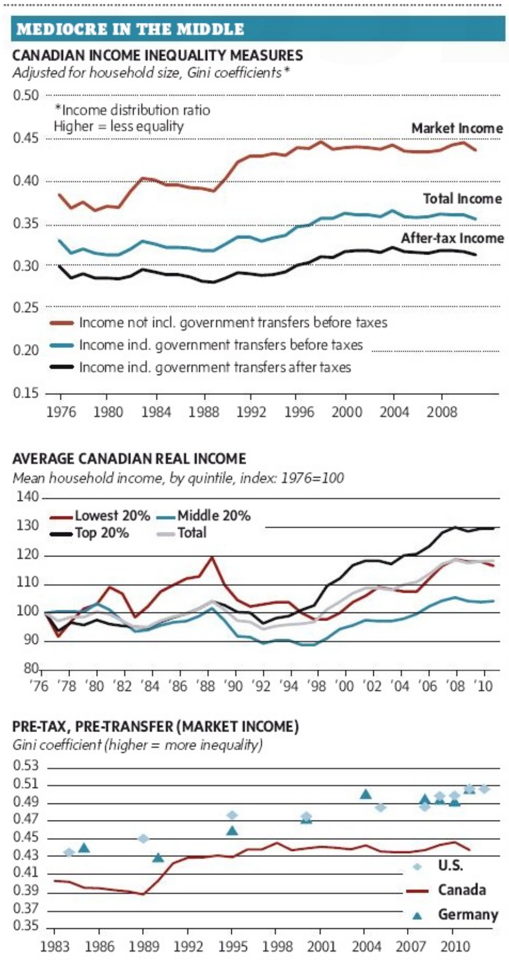
* Baseline = simple average of the country's actual emissions in 2005 and the level expected for 2020 in the absence of international action.

- Etsy, the uber-cool handcraft web site, has a selection of stat-geek gifts you can order from NausicaaDistribution. Just in time!



- and two graphical displays from this week's Globe & Mail, one confusing, one very clear:

MEDIOCRE IN THE MIDDLE



CANADIAN INCOME INEQUALITY MEASURES

*Adjusted for household size, Gini coefficients**

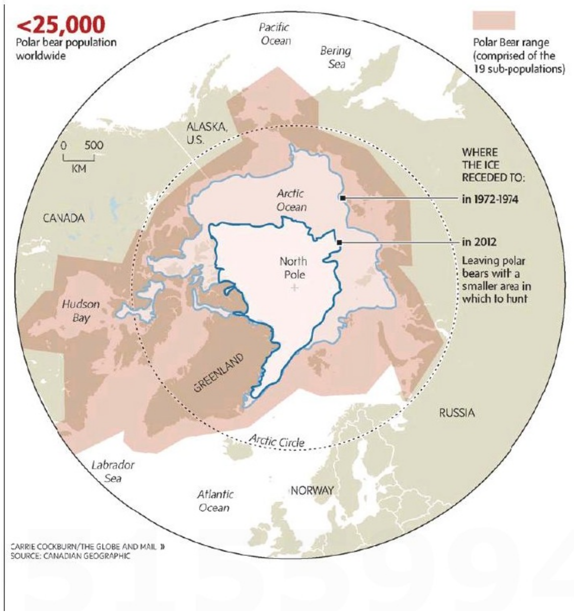
AVERAGE CANADIAN REAL INCOME

Mean household income, by quintile, index: 1976=100

PRE-TAX, PRE-TRANSFER (MARKET INCOME)

Gini coefficient (higher = more inequality)

JOHN SOPNSKI/THE GLOBE AND MAIL » SOURCE: TD ECONOMICS (DATA VIA STATSCAN AND OECD)



<25,000

Polar bear population worldwide

Polar Bear range (comprised of the 19 sub-populations)

WHERE THE ICE RECEDED TO:

in 1972-1974

in 2012

Leaving polar bears with a smaller area in which to hunt

CARRIE COCKBURN/THE GLOBE AND MAIL »
SOURCE: CANADIAN GEOGRAPHIC

Linear Regression

In the class activity on diamond prices, you computed a linear regression of price (in dollars) on weight (in carats: 1 carat = 0.2 grams). Each of you had a different set of data, and all the data was randomly generated based on the results of fitting linear regression to a real data set. The source of the data was 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.

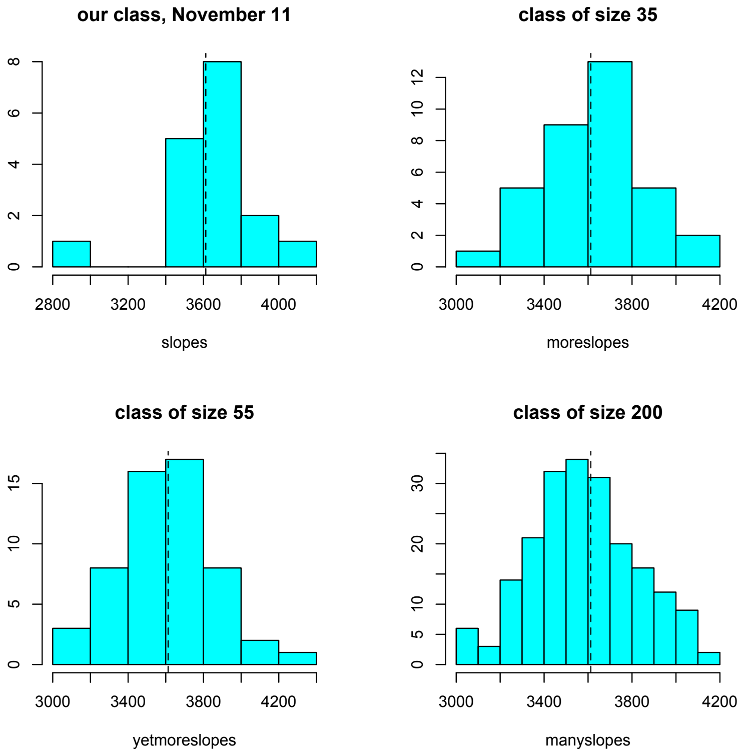
The line you drew is the least squares line for a plot of price (on the y axis) against quality (on the x axis). The regression line has an *intercept* where it crosses the y axis, and a *slope*:

$$y = a + bx.$$

You only need two points to draw a line – one point could be $(0, a)$, but in class I asked you to start at the point $(\bar{x}, \bar{y})$; the least squares line always goes through that point. Then I asked you to go 'over 1' up b, to get another point.

Everybody calculated a different line, because everybody had a unique set of data. Here is the list of slopes:
4030.40 3522.50 3697.97 3752.44 3507.50 3873.90 3949.00 3656.90 3693.90 3689.80 3494.30
2980.10 3477.50 3729.90 3660.30 3476.70 3683.30 (*).

The histogram below (top left) gives a picture. We'd get a larger variety of slopes if we had a bigger class; that's what is shown in the other four pictures.

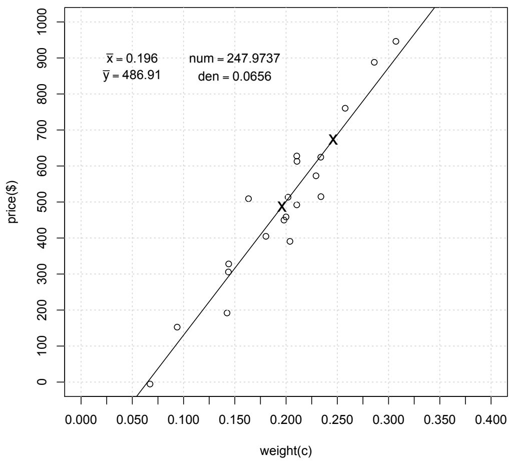


When we compute a linear regression slope, we can also use the same data to estimate the error in our slope. This error is caused by the fact that the data is just a sample from the population, just as it is in a poll.

You computed estimated standard error of your slope, and the list of everyone's standard errors is:

`196.44 236.34 231.35 211.35 219.75 297.75 194.28 231.70 228.71 280.00 202.16 233.16 224.66 342.93 183.04 224.99 256.14.`

The mean of these values is 234.99, and the standard deviation of the list (*) above of slopes above is nearly identical 234.08. In other words, as a class, our estimate of the error was very close to the actual error.



Regression exercise

Your sheet plots the prices of 20 diamonds, of different weights. Your task is to plot the regression line based on your data. I've done a lot of the work for you.

1. The regression line goes through the point $(\bar{x}, \bar{y})$. Mark this point with a cross.

2. The slope of the regression line is

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{num}{den} = \qquad .$$

   I had the computer work out the numerator (top) of this expression and the denominator (bottom). All you have to do is calculate the ratio: call this b.

3. Find a second point on the line by going over 0.05 carats (= 1 grid spacings) and up $b \times 0.05$ dollars, put another cross at the new point and join the two crosses with a line.

4. Work out the *standard error* of the slope using the formula:

$$\text{s.e.}(b) = 58.8/\sqrt{den} = \qquad .$$