Recap

- (Ch 2) Visualizing and summarizing relationships in data
  - Time series data
  - Scatter plots

Today

- (Ch 2) Visualizing and summarizing relationships in data
  - The correlation coefficient
  - Prediction

Correlation coefficient

Given a data set $\{(x, y)\}$ consisting of items $(x_1, y_1), \ldots, (x_N, y_N)$,

1. Standardize the data

$$\hat{x}_i = \frac{x_i - \text{mean}(\{x\})}{\text{std}(\{x\})} \qquad \hat{y}_i = \frac{y_i - \text{mean}(\{y\})}{\text{std}(\{y\})}$$

2. The correlation coefficient is the mean of $\hat{x}_i \hat{y}_i$

$$\text{corr}(\{(x, y)\}) = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \hat{y}_i$$

Correlation, more precisely

In a data set $\{(x, y)\}$ consisting of items $(x_1, y_1), \ldots, (x_N, y_N)$

- we say x and y have **positive correlation** if $\operatorname{corr}(\{(x, y)\}) > 0$
- we say x and y have **negative correlation** if $\operatorname{corr}(\{(x, y)\}) < 0$
- we say x and y have **zero correlation** if $\operatorname{corr}(\{(x, y)\}) = 0$

Properties of the correlation coefficient

- The correlation coefficient is symmetric

$$\mathrm{corr}(\{(x, y)\}) = \mathrm{corr}(\{(y, x)\})$$

- Translating the data does **not** change the correlation coefficient

- Scaling the data may change the sign of the correlation coefficient

$$\mathrm{corr}(\{(ax+b, cy+d)\}) = \mathrm{sign}(ac)\, \mathrm{corr}(\{(x, y)\})$$

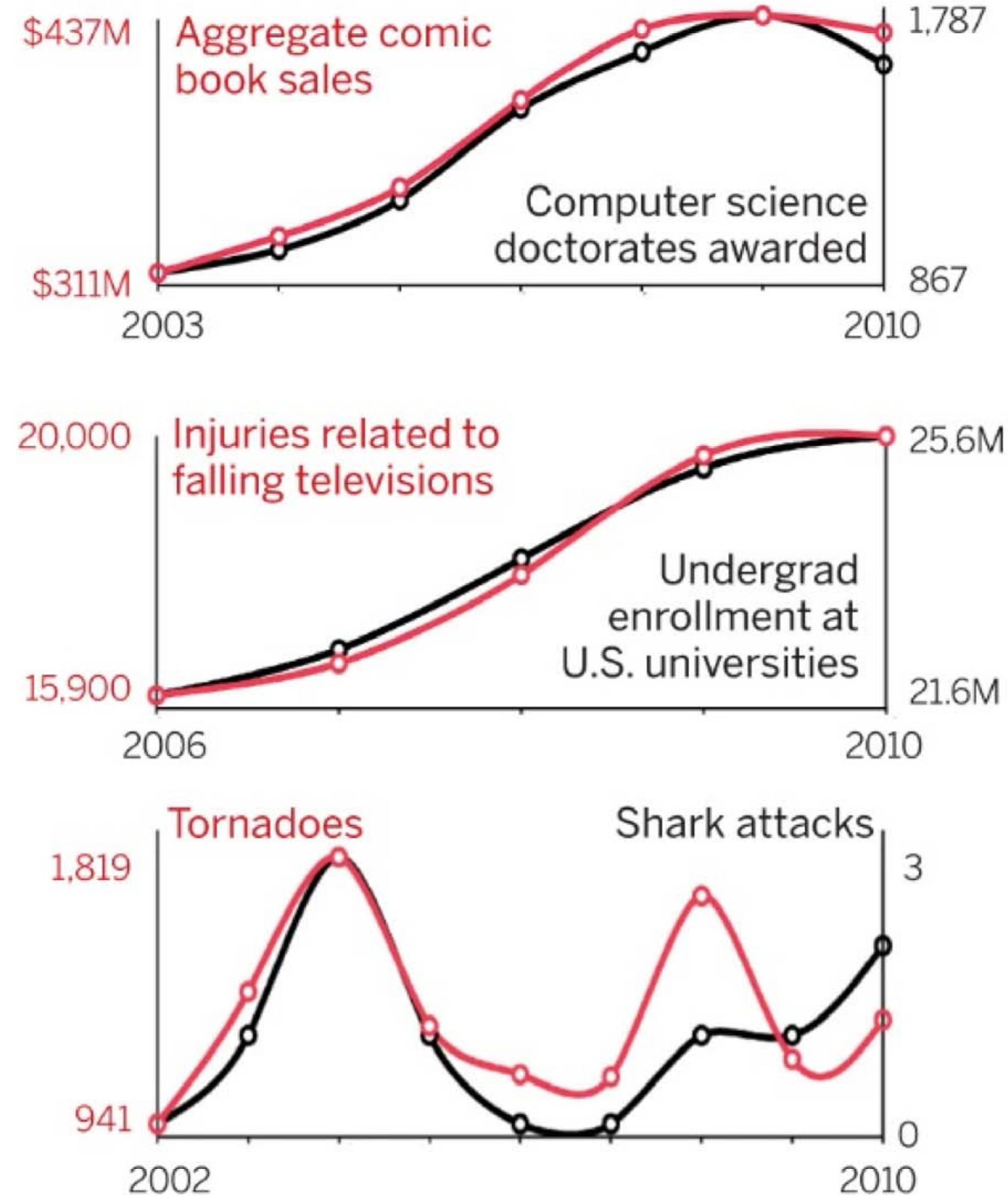
Bounds on the correlation coefficient

The correlation coefficient takes values between -1 and 1 inclusive

$$\text{corr}(\{(x, y)\}) = 1 \text{ if and only if } \widehat{x_i} = \widehat{y_i}$$

$$\text{corr}(\{(x, y)\}) = -1 \text{ if and only if } \widehat{x_i} = -\widehat{y_i}$$

Prediction



From Spurious Correlations by Tyler Vigen

Using correlation to predict

- Given a correlated data set $\{(x, y)\}$,

  we can predict a value y_0^p that goes with a given x_0

- In standard coordinates $\{(\hat{x}, \hat{y})\}$,

  we can predict a value $\hat{y}_0^p$ that goes with a given $\hat{x}_0$

Linear predictor and its error

- We will assume that our predictor is linear

$\hat{y}^p = a\hat{x} + b$, where a & b are constants

- We denote the prediction of $\hat{y}_i$ at each $\hat{x}_i$ in the data set as $\hat{y}_i^p$

$\hat{y}_i^p = a\hat{x}_i + b$

- The error in the prediction of $\hat{y}_i$ is denoted u_i

$u_i = \hat{y}_i - \hat{y}_i^p = \hat{y}_i - a\hat{x}_i - b$

The mean of the prediction error should be 0

$$0 = \text{mean}(\{u_i\})$$

$$= \text{mean}(\{\hat{y}_i - a\hat{x}_i - b\})$$

$$= \underbrace{\text{mean}(\{\hat{y}_i\})}_{0} - a\,\underbrace{\text{mean}(\{\hat{x}_i\})}_{0} - b = -b$$

because $\{\hat{x}\}$ and $\{\hat{y}\}$ are in standard coordinates

$b = 0$

So, the linear predictor becomes

$$\hat{y}^{p} = a\,\hat{x}$$

The variance of the error should be minimal

$$\mathrm{var}(\{u\}) = \mathrm{mean}\left(\left\{\left(u - \underbrace{\mathrm{mean}(\{u\})}_{0}\right)^2\right\}\right)$$

$$= \mathrm{mean}(\{u^2\})$$

$$= \mathrm{mean}\left(\left\{\left(\hat{y} - \hat{y}^p\right)^2\right\}\right)$$

$$= \mathrm{mean}\left(\left\{\left(\hat{y} - a\hat{x}\right)^2\right\}\right)$$

$$= \text{mean}\left(\left\{ (\hat{y})^2 - 2a\hat{x}\hat{y} + a^2(\hat{x})^2 \right\}\right)$$

$$= \text{mean}\left(\left\{ (\hat{y})^2 \right\}\right) - 2a\,\text{mean}\left(\left\{ \hat{x}\hat{y} \right\}\right) + a^2\,\text{mean}\left(\left\{ (\hat{x})^2 \right\}\right)$$

$$= \underbrace{\text{var}\left(\left\{ \hat{y} \right\}\right)}_{1} - 2a\,\underbrace{\text{corr}\left(\left\{ (x,y) \right\}\right)}_{\text{let this be } r} + a^2\,\underbrace{\text{var}\left(\left\{ \hat{x} \right\}\right)}_{1}$$

$$= 1 - 2ar + a^2$$

$$\frac{d}{da}\left(\mathrm{var}(\{u\})\right) = -2r + 2a$$

Setting this to 0 gives $a = r$ minimizes the variance of the error

so $\boxed{\hat{y} = r\hat{x}}$ is the linear predictor

Prediction formulas

- In standard coordinates

$$\hat{y}_0^p = r\hat{x}_0 \text{ where } r = \mathrm{corr}(\{(x, y)\})$$

- In original coordinates

$$\frac{y_0^p - \mathrm{mean}(\{y\})}{\mathrm{std}(\{y\})} = r\,\frac{x_0 - \mathrm{mean}(\{x\})}{\mathrm{std}(\{x\})}$$

Root-mean-square (RMS) prediction error

Recall that

$$\mathrm{var}(\{u\}) = \mathrm{mean}(\{u^2\})$$

$$= 1 - 2ar + a^2$$

$$= 1 - 2r^2 + r^2 \quad \text{because } a = r$$

$$= 1 - r^2$$

So RMS error $= \sqrt{\text{mean}(\{u^2\})}$

$= \sqrt{1-r^2}$

Summary

- We can spot correlation visually in scatter plots
- If the data are strongly correlated (positively or negatively), we can make predictions with small RMS prediction error
- But the ability to make predictions does not guarantee that the predictions are meaningful: correlation does not imply causation!