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 $(x_1, y_1), \dots, (x_N, y_N)$,

- Standardize the data

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

- The correlation coefficient is the mean of $\widehat{x_i}\widehat{y_i}$

$$\mathrm{corr}(\{(x, y)\}) = \frac{1}{N}\sum_{i=1}^{N} \widehat{x_i}\widehat{y_i}$$

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

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

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

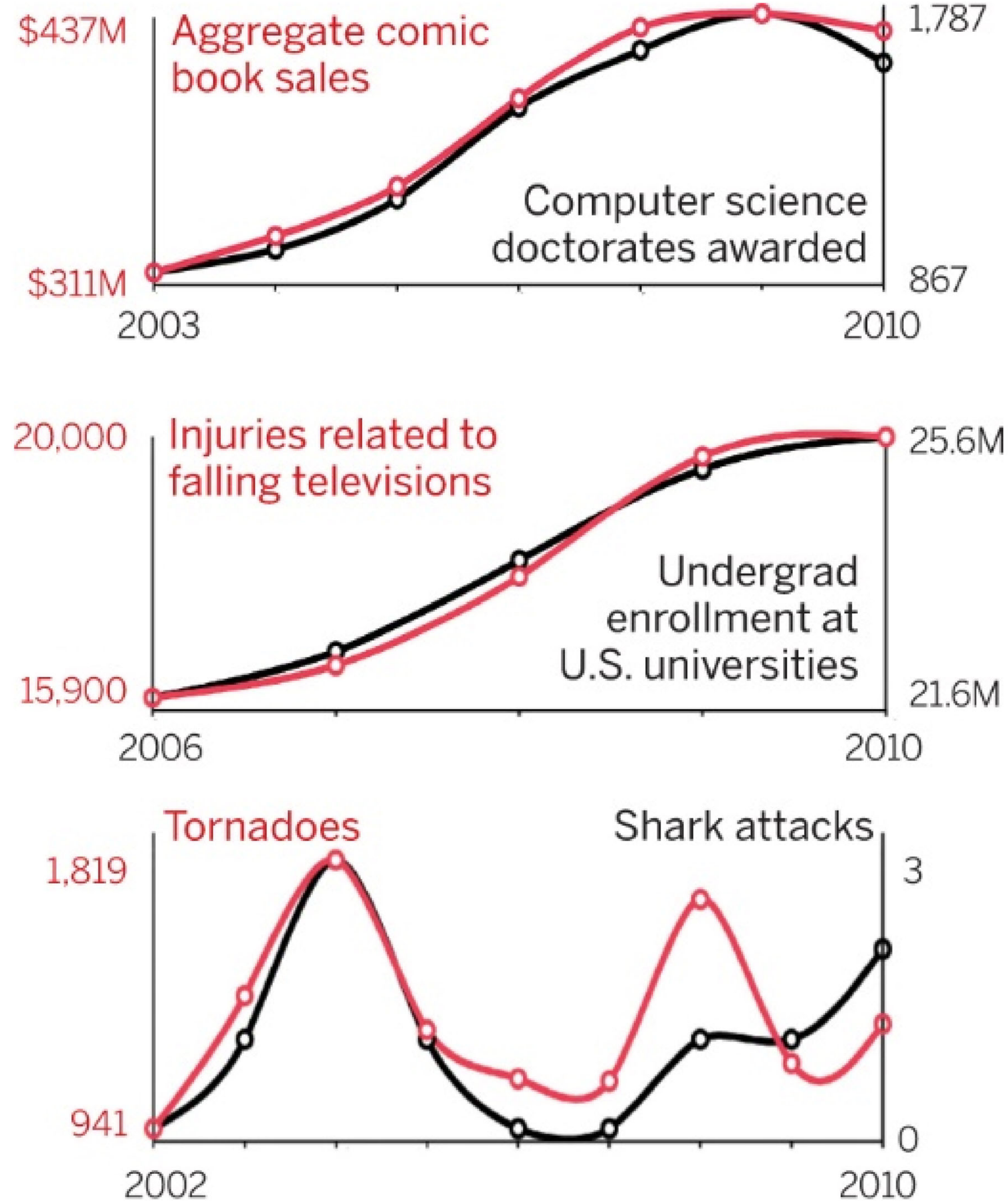
Bounds on the correlation coefficient

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 $\hat{Y}^P = a\hat{X}$

u_i

The variance of the error should be minimal

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

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

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

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

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

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

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

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

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

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

So $\boxed{\hat{y}^p = r\hat{x}}$ is the linear predictor

Prediction formulas

- In standard coordinates

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

- In original coordinates

$$\frac{y_0^p - \text{mean}(\{y\})}{\text{std}(\{y\})} = r\left(\frac{x_0 - \text{mean}(\{x\})}{\text{std}(\{x\})}\right)$$

Root-mean-square (RMS) prediction error

Recall that

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

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

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

$$= 1 - r^2$$

So RMS error $= \sqrt{\mathrm{mean}(\{u^2\})} = \sqrt{1-r^2}$