Today: More on 2-D Continuous Data

Sam Ventura
36-315
Today: More on 2-D and 3-D Continuous Data

Department of Statistics
Carnegie Mellon University

October 10, 2016

correlation $\not\Rightarrow$ causation

Correlation $-1 \le r \le 1$

- $r \approx -1$: strong negative (decreasing) linear relationship
- $r \approx 0$: no linear relationship.
- $r \approx 1$: strong positive (increasing) linear relationship.

2-D Continuous Data – Goals → relationship between the two vars?

Data structure

$\{X_1, X_2\} \in \mathbb{R}^2$ → each variable is continuous

Dataset: n rows, → # obs.
2 columns

Examining variables individually: histograms, boxplots, rugs, density plots, violin plots
- center (mean, median), spread (variance, sd, range, IQR)
skew, modality

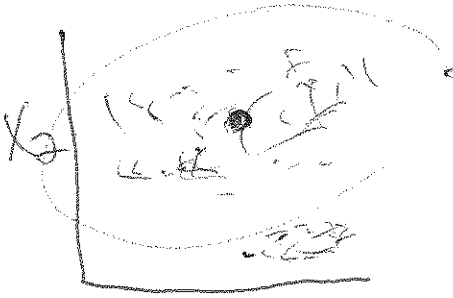
Examining variables together/jointly: "Is there a relationship?" "what is the relationship?"

linear relationship: $y = mx + b$, $X_2 = \hat{\beta}_0 + \hat{\beta}_1 X_1$

nonlinear relationships: $y = f(x)$, $X_2 = f(X_1)$ → any non-linear fn

2-D Continuous Data – Scatterplots

scatterplots: 2-D plot of X_1 and X_2 → mapped to y aesthetic
↳ in the cartesian plane ↳ mapped to x aesthetic



potential outliers

Center → in 2D

Spread → in each direction

What information do we get out of scatterplots?
↳ is there a relationship? how strong? what direction is the linear relationship
↳ Trends (linear or nonlinear)
↳ joint distribution of X_1 and X_2
↳ group structure / clusters of obs
↳ similar to modality in 1-D anx.

2-D Continuous Data – Trends

Linear trends: linear regression (simple)
uses "least squares" to find the "best" linear fit between X_1 and X_2. Trying to estimate β_0 and β_1 in this equation: $X_2 = \beta_0 + \beta_1 X_1$ $(y = mx + b)$

short version: minimizes sum of squared errors from points to the regression line

Non-linear trends: "smoothers" splines, lowess regression/smoothers
↳ fit non-linear/non-parametric, "local" regression to X_1, X_2
Uses a "moving window" to estimate the best fit at each part of the data

potential issues: pick bandwidth
over fitting, especially to outliers