MATH 205: Statistical methods

Lecture 5: Using correlation to predict

Reminders

- First homework was posted on the course webpage.
- Due: next Friday before lecture
- You can submit the homework online or in person (paper)

Tentative schedule

| Date	Theme/Topic	Labs	Assignments
Aug 31	Syllabus		
Sep 2—9	Chapter 1: Describing dataset	Section 2: Handling data	
Sep 12—16	Chapter 2: Looking at Relationships	Section 3: Univariate data	
Sep 19—23	Chapter 3: Basic Ideas in Probability	Section 4: Bivariate Data	Homework 1 (due 09/23)
Sep 26—30	Chapters 3-4	Section 4: Correlation	
Oct 3—7	Chapter 4: Random variables and expectations	Section 6: Random data	Homework 2 (due 10/07)
Oct 10—14	Chapter 5: Useful distributions	Section 7: The central limit theorem	
Oct 17—21	Chapter 6: Samples and populations	Section 9: Confidence interval estimation	Homework 3 (due 10/21)
Oct 24—28	Review Midterm exam		Midterm: Oct 28 (lecture) Oct 24-26 (labs)
Oct 31—Nov 4	Chapter 7: The significance of evidence	Section 10: Hypothesis testing	
Nov 7—11	Goodness of Fit	Section 12: Goodness of Fit	Homework 4 (due 11/11)
Nov 14—18	Linear Regression	Section 13: Linear regression	
Nov 21–25	Thanksgiving break		
Nov 28 —Dec 2	One-Way Analysis of Variance	Section 15: Analysis of variance	Homework 5 (due 12/02)
Dec 5—7	Selected topics + Review		
Exam week			

Chapter 2: Looking at relationship

Plotting 2D data

- We take a dataset, choose two different entries, and extract the corresponding elements from each tuple
- The result is a dataset consisting of 2-tuples, and we think of this as a two dimensional dataset
- Goal: to plot this dataset in a way that reveals relationships
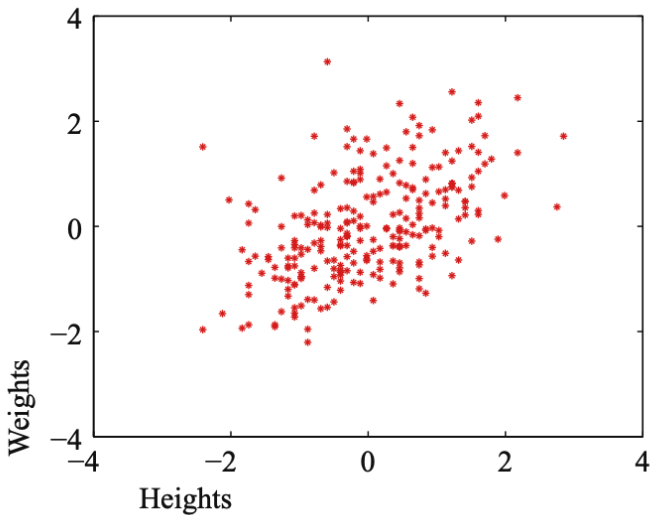
- categorical vs categorical: create a richer set of categories
- categorical vs continuous: comparative box plots
- continuous vs continuous: scatter plots
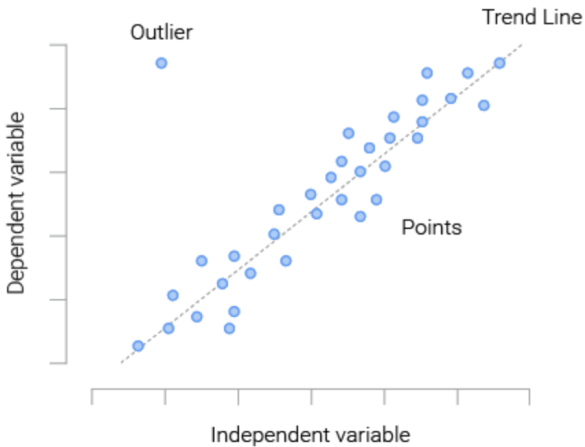
Standard coordinates

Question: when $\hat{x}$ increases, does $\hat{y}$ tend to increase, decrease, or stay the same?

- Positive correlation: larger $\hat{x}$ values tend to appear with larger $\hat{y}$ values
- Negative correlation: larger $\hat{x}$ values tend to appear with smaller $\hat{y}$ values
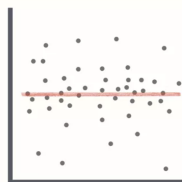- Zero correlation: no relationship

Correlations

Positive Correlation

Negative Correlation

No Correlation

Definition 2.1 (Correlation Coefficient)   Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \dfrac{(x_i - \mathsf{mean}(\{x\}))}{\mathsf{std}(x)}$, $\hat{y}_i = \dfrac{(y_i - \mathsf{mean}(\{y\}))}{\mathsf{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\mathsf{corr}(\{(x, y)\}) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

Correlation coefficient

- correlation is a measure of our ability to predict one value from another
- correlation coefficient takes values between -1 and 1
- If the correlation coefficient is close to 1 or -1, then we are likely to predict very well.

Correlation coefficient: properties

Useful Facts 2.1 (Properties of the Correlation Coefficient)

- The correlation coefficient is symmetric (it doesn't depend on the order of its arguments), so

$$\text{corr}\left(\{(x, y)\}\right) = \text{corr}\left(\{(y, x)\}\right)$$

- The value of the correlation coefficient is not changed by translating the data. Scaling the data can change the sign, but not the absolute value. For constants $a \neq 0$, b, $c \neq 0$, d we have

$$\text{corr}\left(\{(ax + b, cx + d)\}\right) = \text{sign}(ab)\text{corr}\left(\{(x, y)\}\right)$$

- If $\hat{y}$ tends to be large (resp. small) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be positive.
- If $\hat{y}$ tends to be small (resp. large) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be negative.
- If $\hat{y}$ doesn't depend on $\hat{x}$, then the correlation coefficient is zero (or close to zero).
- The largest possible value is 1, which happens when $\hat{x} = \hat{y}$.
- The smallest possible value is -1, which happens when $\hat{x} = -\hat{y}$.

Confusion caused by correlation

Correlation does not imply causation

- When two variables are correlated, they change together. This means that one can make a reasonable prediction of one from the other.
- However, correlation does not mean that changing one variable causes the other to change (sometimes known as causation).
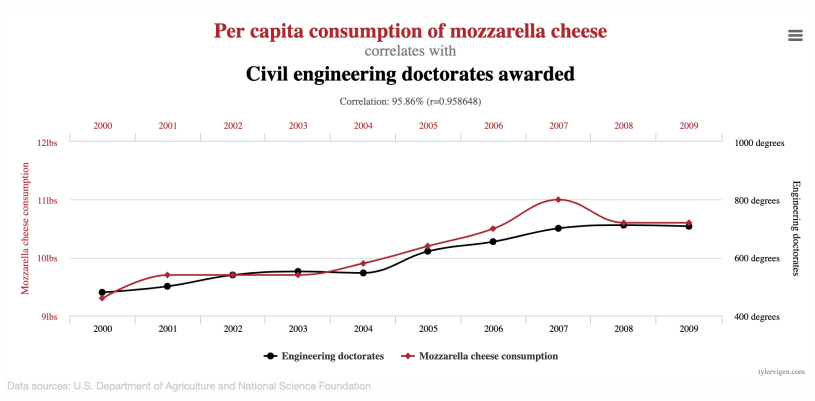
Variables could be correlated for a variety of reasons

Background (latent) variable:

- In children, shoe size is correlated with reading skills
- This doesn't mean that making your feet grow will make you read faster, or that you can make your feet shrink by forgetting how to read
- Latent variable: age. Young children tend to have small feet, and tend to have weaker reading skills

Variables could be correlated for a variety of reasons

Random chances:

Per capita consumption of mozzarella cheese
correlates with
Civil engineering doctorates awarded

Correlation: 95.86% (r=0.958648)

Data sources: U.S. Department of Agriculture and National Science Foundation

tylervigen.com

Using correlation to predict

- Assume that we have a two-dimensional datasets of N points:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$$

  $\rightarrow$ we can compute the correlation coefficient r

- Assume that r is close to 1, so we are confidence that we can predict y from x
- Assume that we have a new data point $(x_0, ?)$
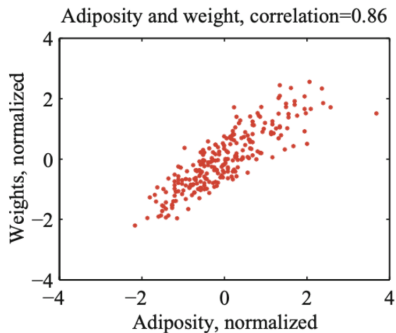- Question: How do we predict this unknown value '?'

$$\hat{x}_i = \frac{1}{\text{std}(x)}(x_i - \text{mean}(\{x\}))$$

$$\hat{y}_i = \frac{1}{\text{std}(y)}(y_i - \text{mean}(\{y\}))$$

$$\hat{x}_0 = \frac{1}{\text{std}(x)}(x_0 - \text{mean}(\{x\})).$$

Idea: If we can predict the corresponding value $\hat{y}_0$, then we can transform back to the original coordinates and make prediction

Adiposity and weight, correlation=0.86

Idea: Maybe we could use a linear function to predict $\hat{y}$ from $\hat{x}$?

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

and find a, b such that $\hat{y}_i - \hat{y}_i^p \approx 0$

- Denote

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

- We want $u_i \approx 0$
- One way to do that is find a, b to ensure that

    mean$(\{u\}) = 0$,   and   std$(\{u\})$ as small as possible

$$\text{mean}\left(\{u\}\right) = \text{mean}\left(\{\hat{y} - \hat{y}^p\}\right)$$
$$= \text{mean}\left(\{\hat{y}\}\right) - \text{mean}\left(\{a\hat{x}_i + b\}\right)$$
$$= \text{mean}\left(\{\hat{y}\}\right) - a\text{mean}\left(\{\hat{x}\}\right) + b$$
$$= 0 - a0 + b$$
$$= 0.$$

We deduce that b should be 0.

$$\text{var}\left(\{u\}\right) = \text{var}\left(\{\hat{y} - \hat{y}^p\}\right)$$

$$= \text{mean}\left(\{(\hat{y} - a\hat{x})^2\}\right) \quad \text{because mean}\left(\{u\}\right) = 0$$

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

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

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

For a fixed value of r, the optimal value for a is $a = r$ and the corresponding value for $var(\{u\})$ is $1 - r^2$.

Procedure 2.1 (Predicting a Value Using Correlation) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. Assume we have an x value x_0 for which we want to give the best prediction of a y value, based on this data. The following procedure will produce a prediction:

- Transform the data set into standard coordinates, to get

$$\hat{x}_i = \frac{1}{\text{std}(x)}(x_i - \text{mean}(\{x\}))$$

$$\hat{y}_i = \frac{1}{\text{std}(y)}(y_i - \text{mean}(\{y\}))$$

$$\hat{x}_0 = \frac{1}{\text{std}(x)}(x_0 - \text{mean}(\{x\})).$$

- Compute the correlation

$$r = \text{corr}(\{(x, y)\}) = \text{mean}(\{\hat{x}\hat{y}\}).$$

- Predict $\hat{y}_0 = r\hat{x}_0$.
- Transform this prediction into the original coordinate system, to get

$$y_0 = \text{std}(y)r\hat{x}_0 + \text{mean}(\{y\})$$

By using the prediction procedure above, we have the error in prediction is

$$var(\{u\}) = 1 - r^2$$

Thus, the closer r^2 to 1, the better the prediction.