Summary of Lecture 2

We have covered three topics in Lecture 2.

1. Linear functions and combinations of variables

   Sections 2.4.3 and 2.4.4 of "OpenIntro statistics" discuss closely related topic of linear combinations of random variables.
   In order to fully understand this, you will need to read through Chapter 2. We will cover random variables in Lecture 3, 4 and 5, so reading Chapter 2 will prepare you for the next week's lecture.

2. Introduction to linear regression

   Sections 7.1 and 7.2 "OpenIntro statistics" provide a good introduction to linear regression.

3. Clustering

   Clustering is not covered in "OpenIntro statistics."
   Sections 8.1 and 8.2 of http://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf provide a very lucid introduction.

Variable y is a linear function of variable x if

$$y = c_0 + c_1 \cdot x,$$

where c_0 is an intercept and c_1 is a slope.

Given the sample mean and variance of x, we can obtain the sample mean of y as

$$\bar{y} = c_0 + c_1 \cdot \bar{x},$$

the sample variance as

$$s_y^2 = c_1^2 \cdot s_x^2$$

and the sample standard deviation as $s_y = |c_1| \cdot s_x$.

The effect of c_1 is that it affects proportionally the sample mean and standard deviation, while c_0 just changes the mean.

If

$$y = c_0 + c_1 x_1 + c_2 x_2,$$

then y is a linear combination of 2 different variables x_1 , x_2. Given the sample means and variances of x_1 , x_2, we have $\overline{y} = c_0 + c_1 \overline{x}_1 + c_2 \overline{x}_2$ and $s_y^2 = c_1^2 s_{x_1}^2 + c_2^2 s_{x_2}^2 + 2 c_1 c_2 s_{x_1 x_2}$.

When $y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3$, then

$$\overline{y} = c_0 + c_1 \overline{x}_1 + c_2 \overline{x}_2 + c_3 \overline{x}_3$$

and

$$s_y^2 = c_1^2 s_{x_1}^2 + c_2^2 s_{x_2}^2 + c_3^2 s_{x_3}^2 + 2 \left[c_1 c_2 s_{x_1 x_2} + c_1 c_3 s_{x_1 x_3} + c_2 c_3 s_{x_2 x_3} \right].$$

These formulas are useful in determining the average return on a portfolio, as well as its risk. We have

$$R_p = \sum_{i=1}^{m} w_i x_i$$

that is, a portfolio is a linear combination of individual asset returns.

Linear regression is a statistical tool that allows us to find the line that fits data the best. (Univariate) linear regression helps us find the **linear relationship** or **linear function** between two variables

$$y = c_0 + c_1 \cdot x.$$

The variable c_0 is the **intercept**. The variable c_1 is a number called the **slope**. Using the regression line we can predict y using the new value x.

The slope coefficient can be obtained using the formula: slope $= \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x} \cdot r_{xy}$. The slope formula takes the covariance and "standardizes" it so that its units are (units of y)/(units of x).

Intercept can be obtained using the formula: intercept $= \overline{y} - \text{slope} \cdot \overline{x}$. The intercept formula will make the regression line pass through the point $(\overline{x}, \overline{y})$.

The formulas for the slope and intercept just use the sample mean, sample covariance, and sample variance.

Limitations of regression:

1. Regression only finds a linear relationship between y and x. That is, it fill fit a line, even though there may be a nonlinear relationship between y and x.
2. Just because we can fit a line through data, it does not mean that changes in x cause changes in y. That is, correlation does not imply causation.
3. We can use regression only to predict in the region where we have observations x. That is, we cannot predict y for values of x that are far from the observed x values.

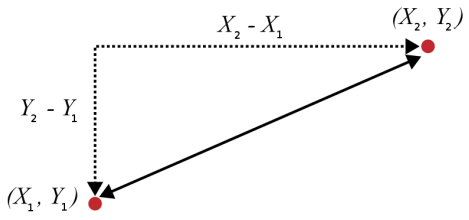
Clustering is a tool that allows us to find groups of similar observations.

In order to group observations, we need a similarity measure between observations. We often compute distance between observations rather than similarity. Observations that are far from each other are less similar to each other. The Euclidean distance is commonly used and is given as

$$\sqrt{(A_1 - A_2)^2 + (B_1 - B_2)^2 + \ldots + (Z_1 - Z_2)^2}.$$

In a special case when there are two varialbes, the distance between two observations can be computed as

$$\sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}.$$



Clustering is an exploratory tool. As such, the number of groups is determined by common sense, domain expertize or business demands.

The k-means clustering algorithm finds k groups and assigns each observation to the closest group. The groups are identified so that the total distance between each observation and the closest group center is minimized.

Each group of observations is represented by the group's center.