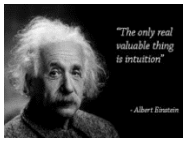
Principal Components Regression (PCR)



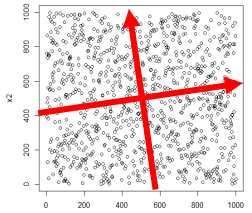
Principal Components (PCs): Intuition

- If you have **P variables** and try every possible linear combination of the variables, you can then identify the one with the **highest variance**
- This linear combination where the data has more variance is called the **"First Principal Component"**
- Then evaluate at all the linear combinations that are **perpendicular** to this component and find the one with the highest variance is called the **"Second Principal Component"**; and so on.
- A data set with **P variables** has exactly **P PCs**
- **Two** very nice **properties** of these PC's are that:
  - They are all **perpendicular** (i.e., **"orthogonal")** with each other, thus they are **independent** with 0 correlation → **no multicollinearity**)
  - They are **sorted** from **highest to lowest variance** dimensions

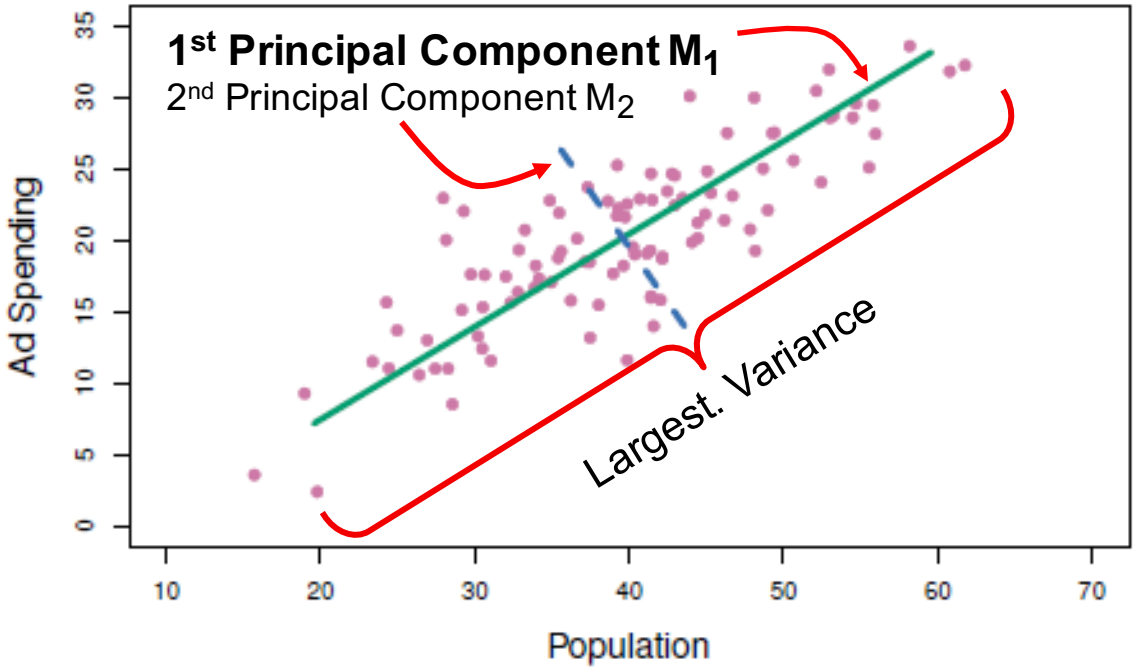
Principal Components: Illustration

This can be **illustrated** with **2 variables** like population size and advertising expenditures (as predictors of sales), resulting in **2 Principal Components**

PC's not helpful here



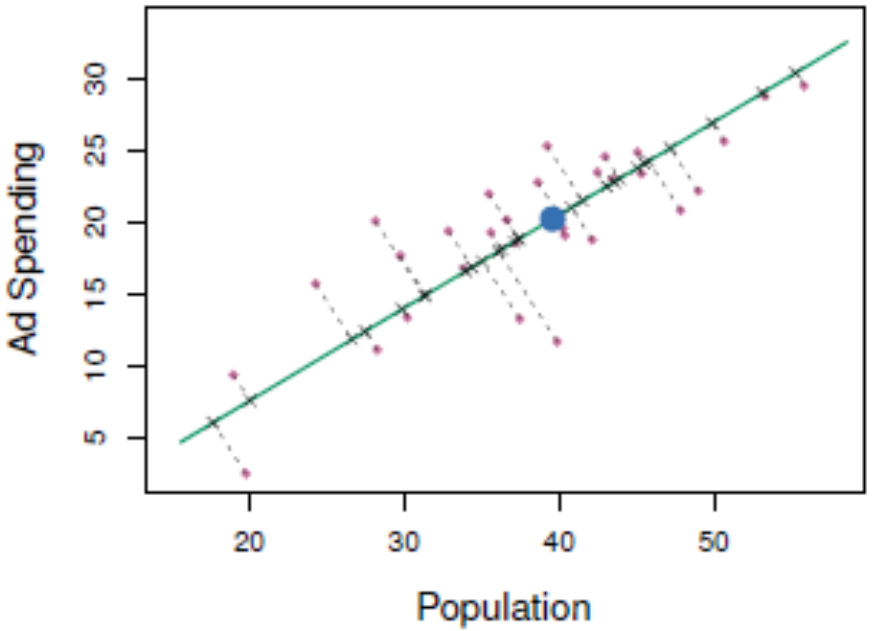
Think of Principal component as **rotating** the **axes** into the **highest variance** direction and then moving the **origin** to the mean (i.e., **centering**) of the variables involved.

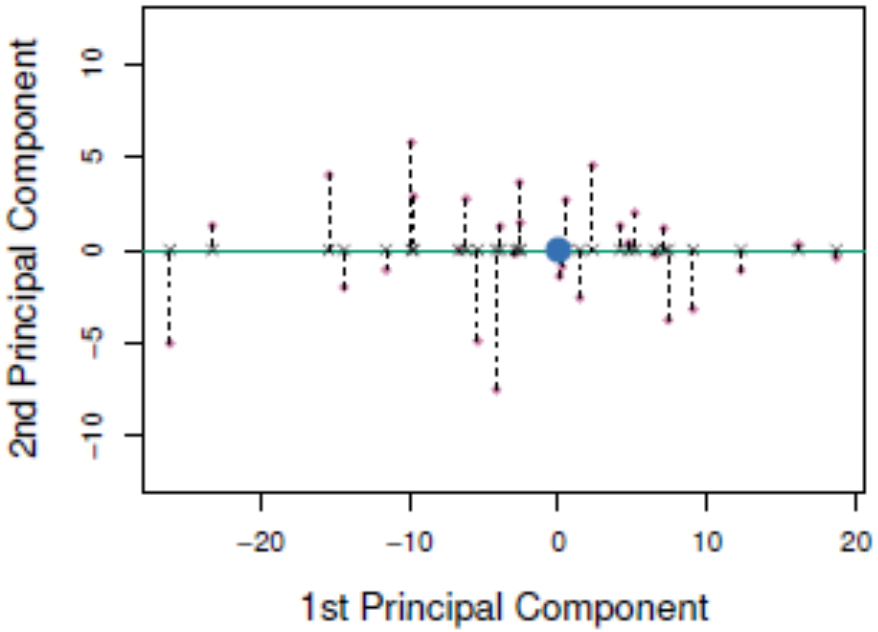


It is clear from this plot that **Ad Spending** and **Population** are highly **correlated**. But Principal Components **M_1** and **M_2** are **not**

Orthogonality

Once we rotate the 2 PC's it becomes clear that the two perpendicular components have 0 correlation – i.e., they are **"independent"** or **"orthogonal"**





KOGOD SCHOOL
of
BUSINESS