Unsupervised Learning : PCA

Swati Mishra

Applications of Machine Learning (4AL3)

Fall 2024

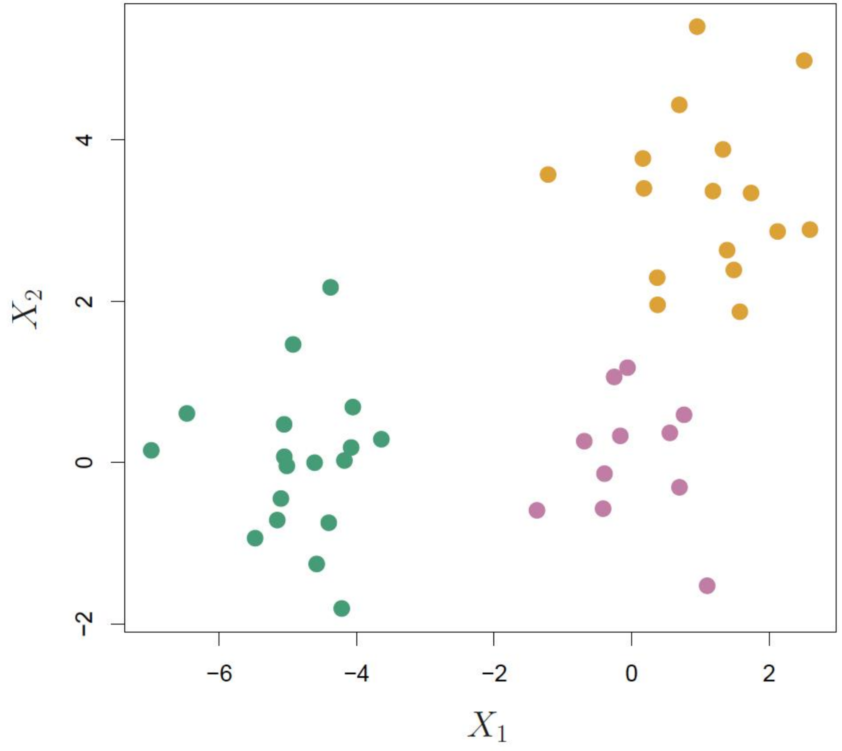
McMaster University

ENGINEERING

Review

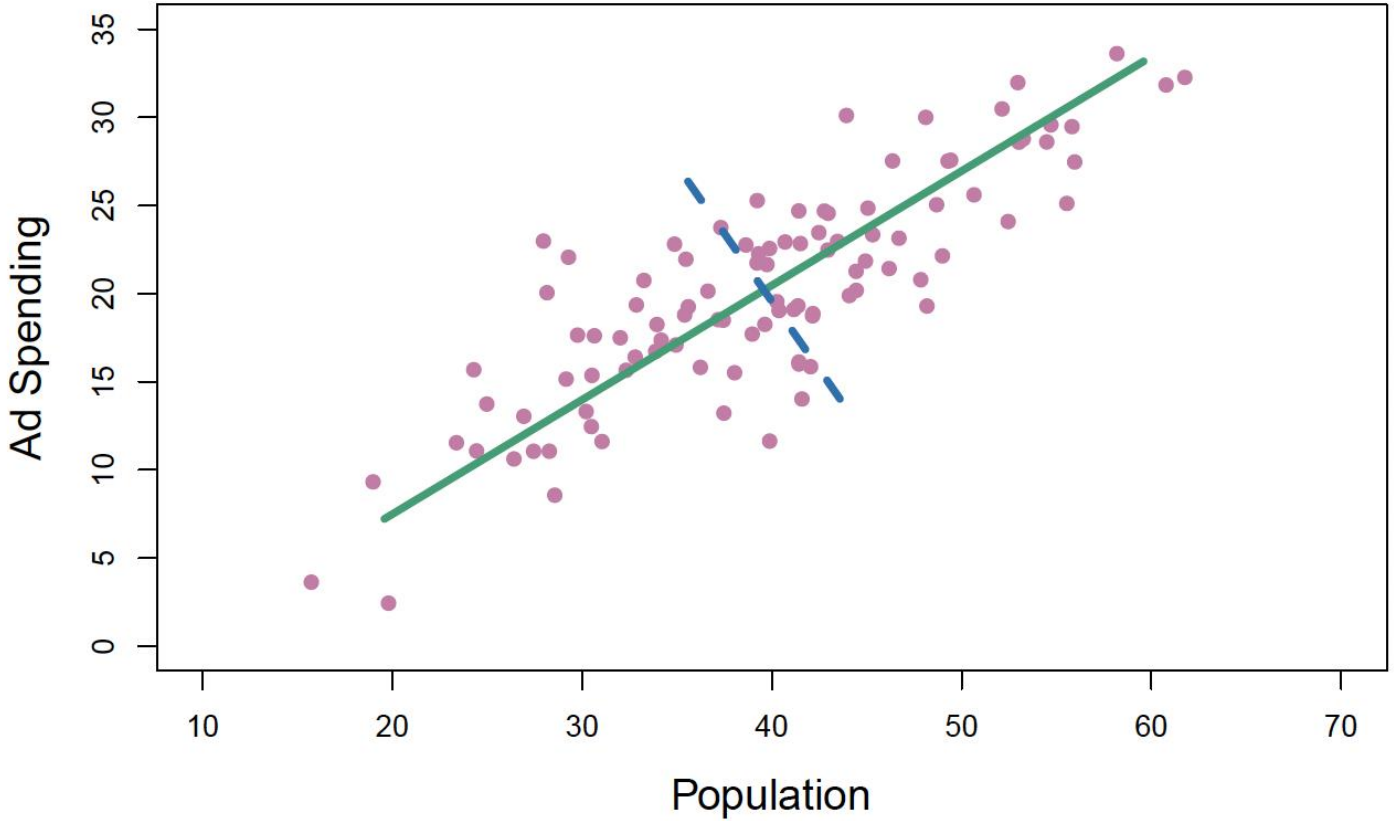
- Hierarchical Clustering
- Dendrograms
- Linkage Techniques
- Introduction to PCA, Projections

Projections



How do we project a sample in this space?

Dimensionality Reduction

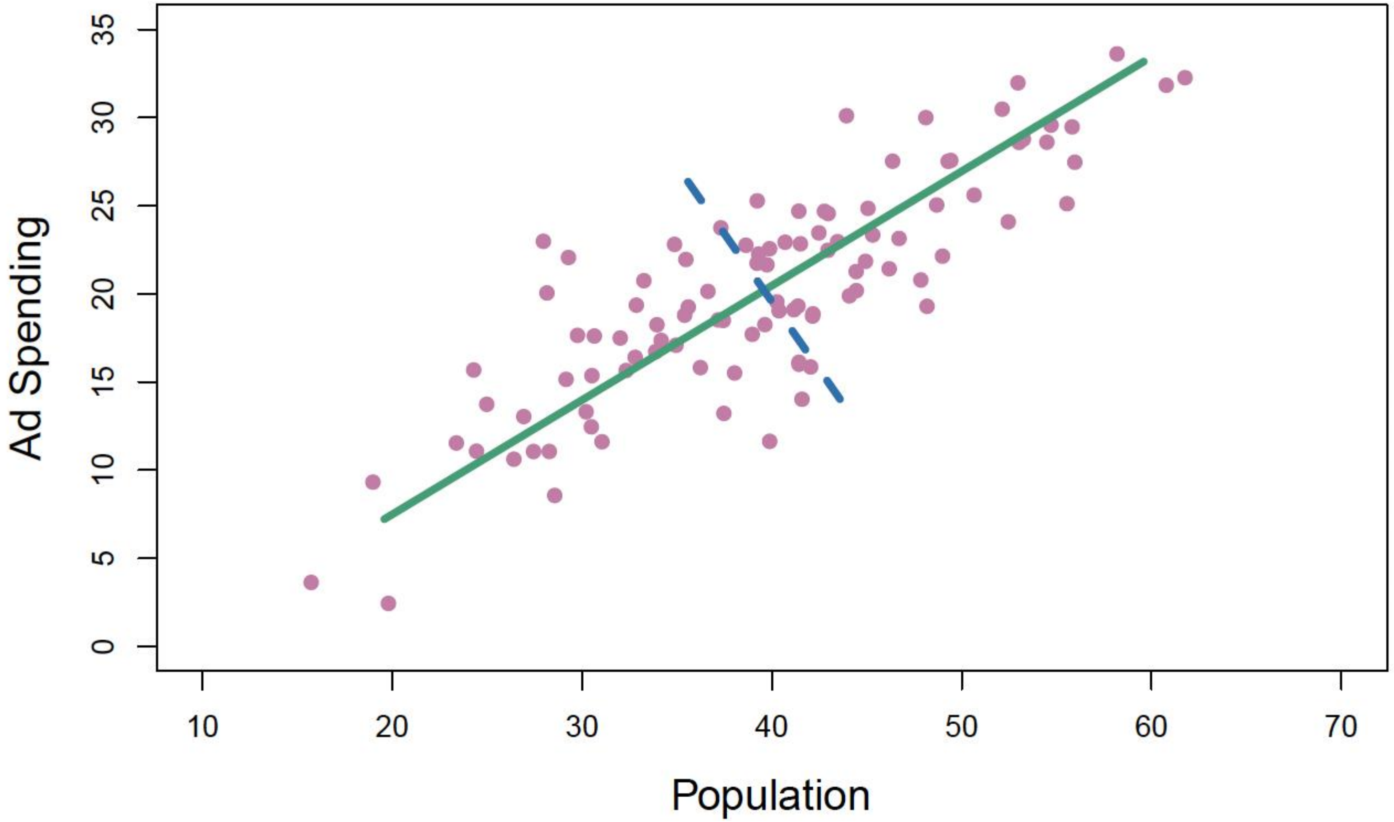


Let Z_1 be a linear combination of p

$$Z_i = \emptyset_{11} X_1 + \emptyset_{21} X_2 + \ldots + \emptyset_{p1} X_p$$

McMaster University

Dimensionality Reduction



Let Z_l be a linear combination of p

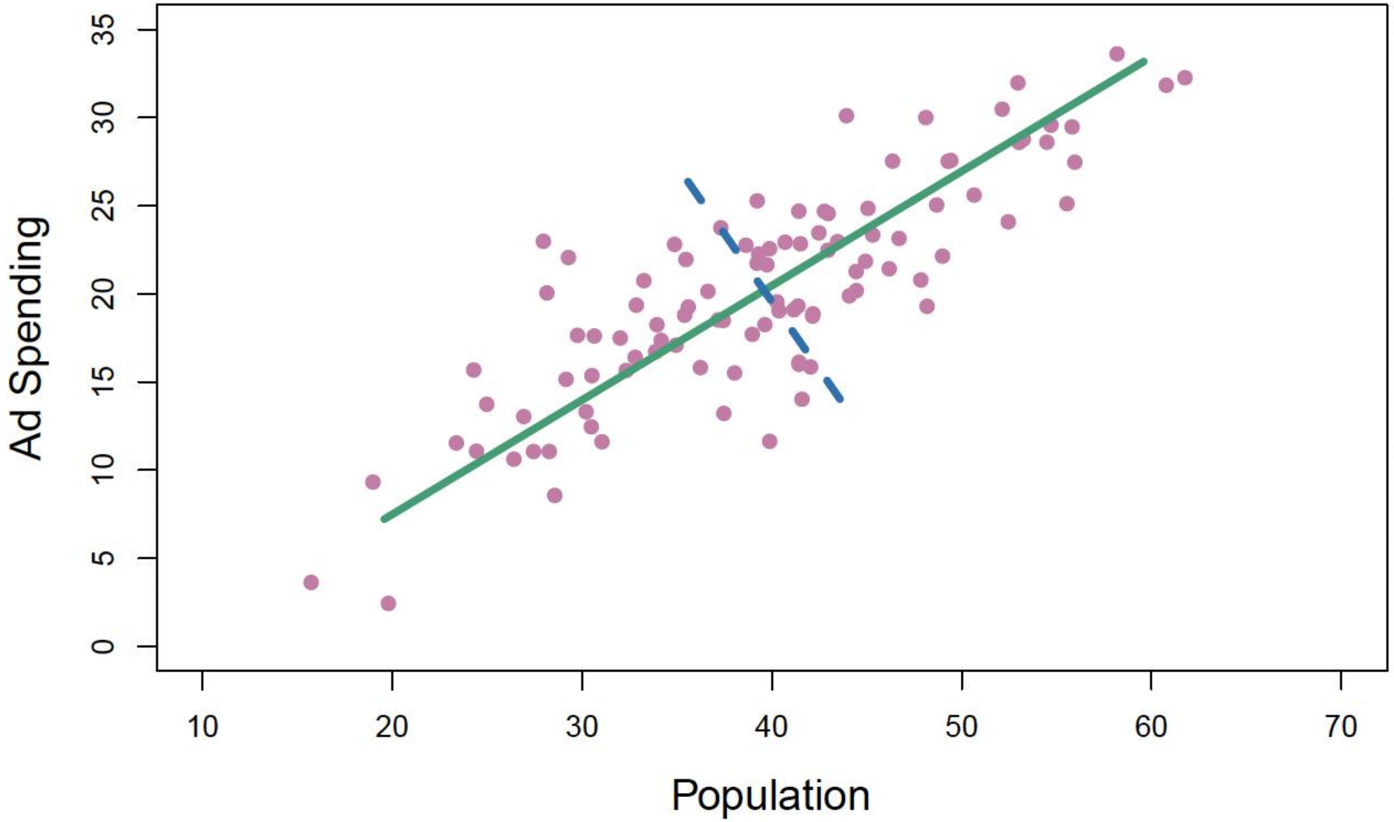
$$Z_i = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

This can be written as:

$$Z_m = \sum_{j=1}^{p} \emptyset_{jm}X_j$$

McMaster University

Dimensionality Reduction



Let Z_{l} be a linear combination of p

$$Z_i = \emptyset_{11}X_1 + \emptyset_{21}\,X_2 + \ldots + \emptyset_{p1}X_p$$

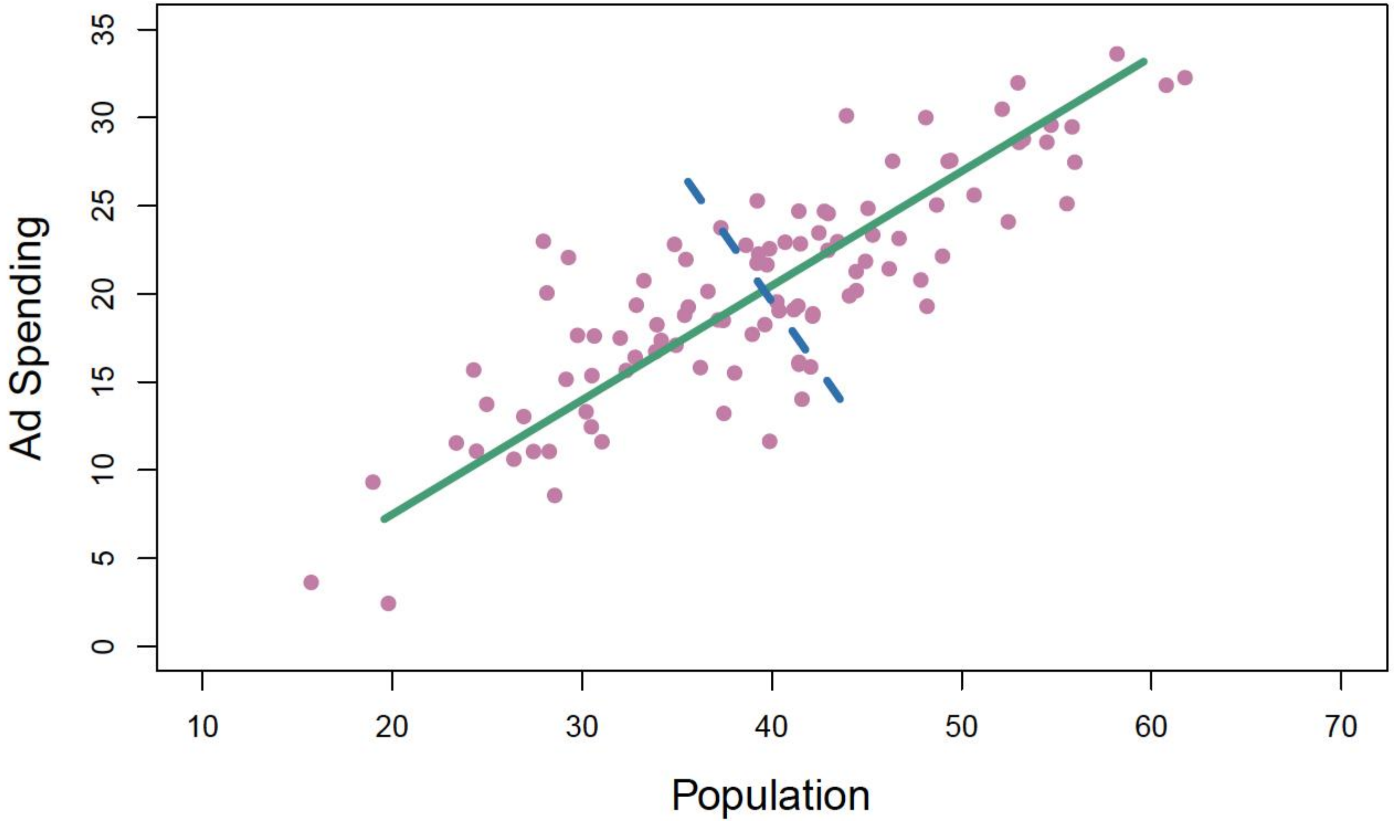
This can be written as:

$$Z_m = \sum_{j=1}^{p} \emptyset_{jm}X_j$$

This can also be written as:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_m + \epsilon_i$$

Dimensionality Reduction



Let Z_l be a linear combination of p

$$Z_i = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

This can be written as:

$$Z_m = \sum_{j=1}^{p} \emptyset_{jm}X_j$$

This can also be written as:

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_m + \epsilon_i$$

The above equation can be solved using ordinary least squares

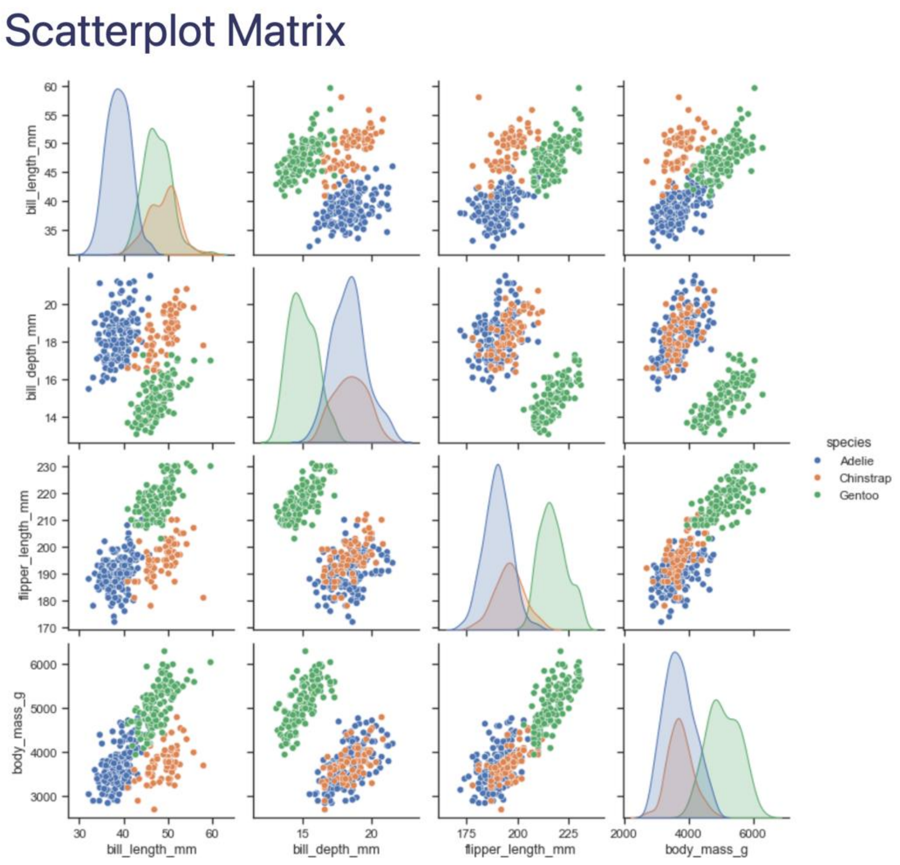
Principal Component Analysis

- Suppose that we wish to visualize n observations with measurements on a set of p features.

- Not all of p dimensions are equally interesting.

- PCA finds a low-dimensional representation of a data set that contains as much as possible of the variation.

- PCA seeks a small number of dimensions that are as interesting as possible.



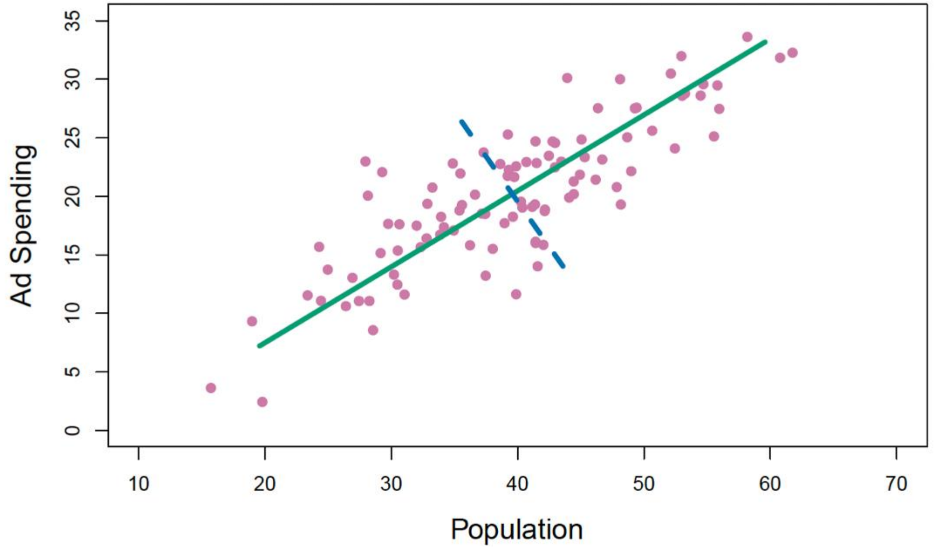
McMaster University

Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA



Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

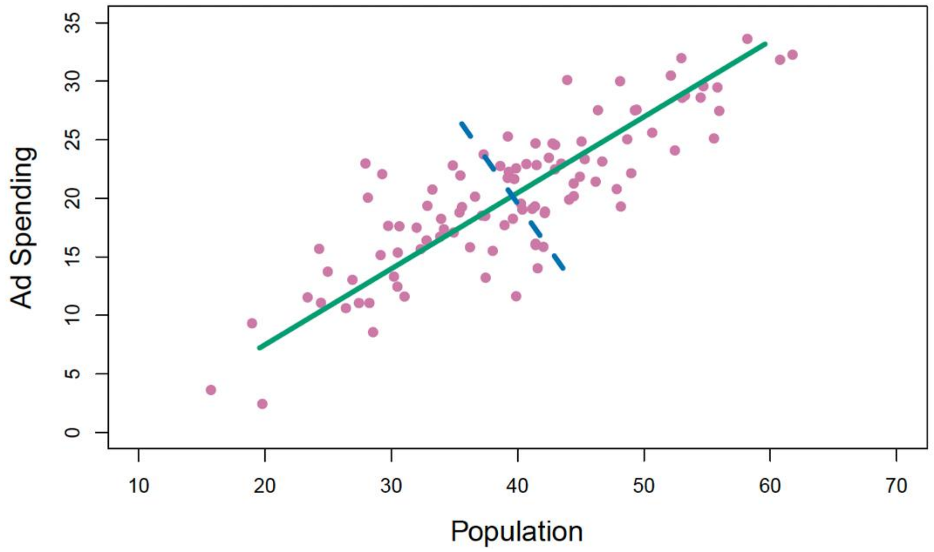
$$Z = \emptyset_{11} X_1 + \emptyset_{21} X_2 + \ldots + \emptyset_{p1} X_p$$

  - $\emptyset_{11}$ = loadings of the PCA

$$Z_m = \sum_{j=1}^{p} \emptyset_{jm} X_j$$

This can be written as: $y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_m + \epsilon_i$

  - The above equation can be solved using ordinary least squares



Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

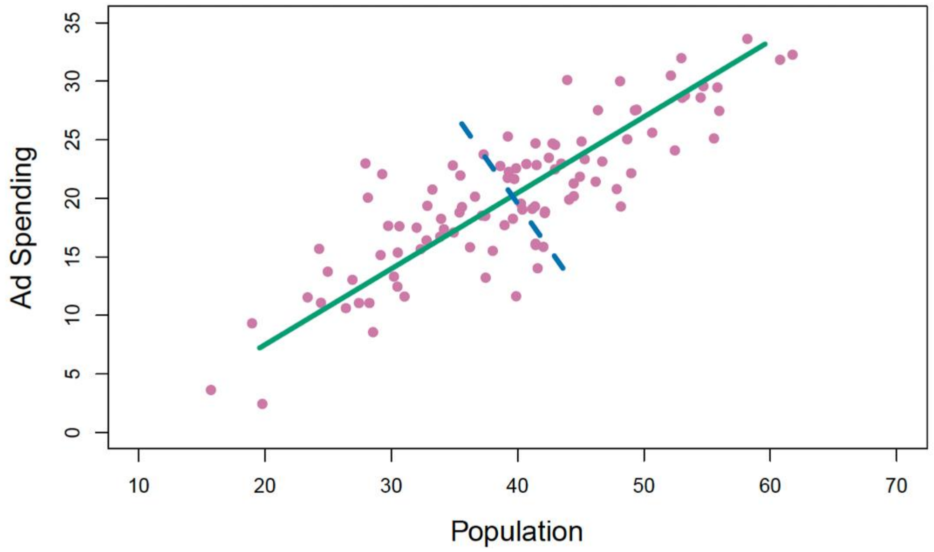
$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA

$$Z_m = \sum_{j=1}^{p} \emptyset_{jm}X_j$$

This can be written as: $y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_m + \epsilon_i$

  - The above equation can be solved using ordinary least squares



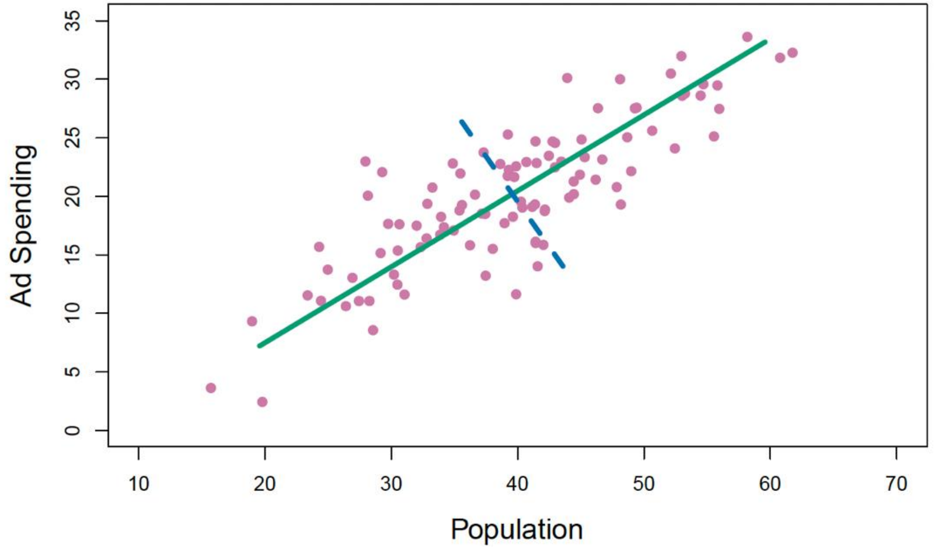
Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}\, X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11} =$ loadings of the PCA

Let’s assume: $\emptyset_{11} = 0.839\,,\, \emptyset_{21} = 0.544$



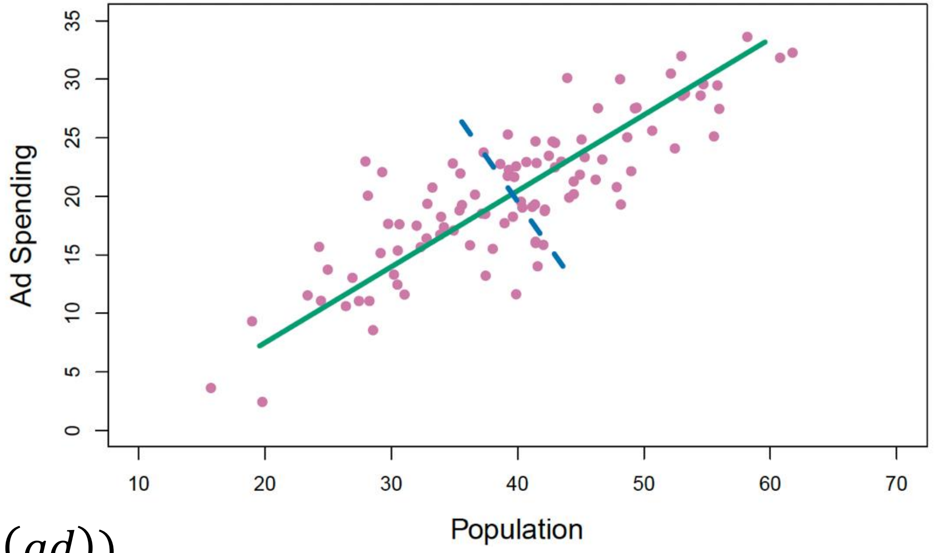
Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}\, X_2 + \ \ldots + \ \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA

Let's assume: $\emptyset_{11} \ = 0.839\, , \emptyset_{21} \ = 0.544$



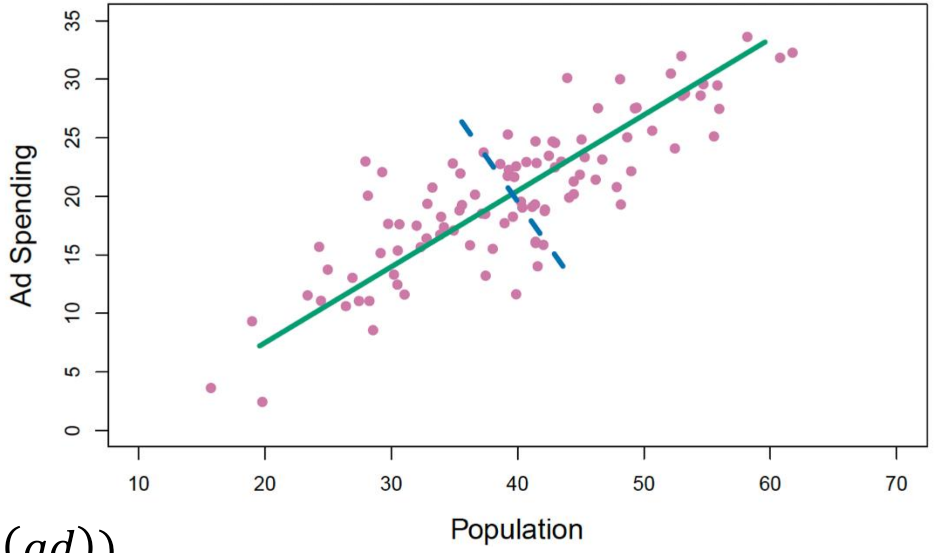
$$Z_1 = 0.839\ \times \left(pop\ - mean(pop)\right) + 0.544\ \times (ad\ - mean(ad))$$

Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA



Let's assume: $\emptyset_{11} = 0.839$, $\emptyset_{21} = 0.544$

$$Z_1 = 0.839 \times (pop - mean(pop)) + 0.544 \times (ad - mean(ad))$$

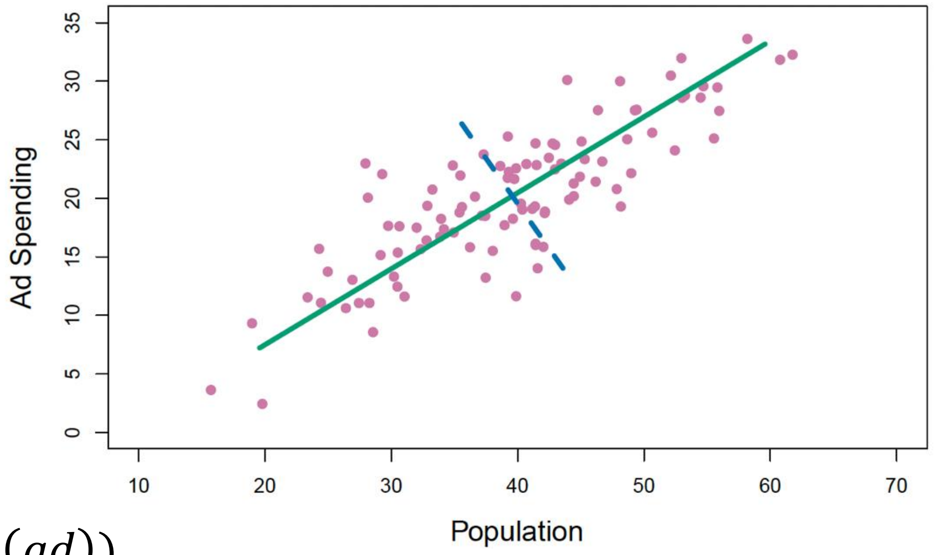
McMaster University

Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA



Let's assume: $\emptyset_{11} = 0.839$, $\emptyset_{21} = 0.544$

Then, $\emptyset_{11} \times \emptyset_{11} + \emptyset_{21} \times \emptyset_{21} = 1$

$$Z_1 = 0.839 \times (pop - mean(pop)) + 0.544 \times (ad - mean(ad))$$

McMaster University

Principal Component Analysis

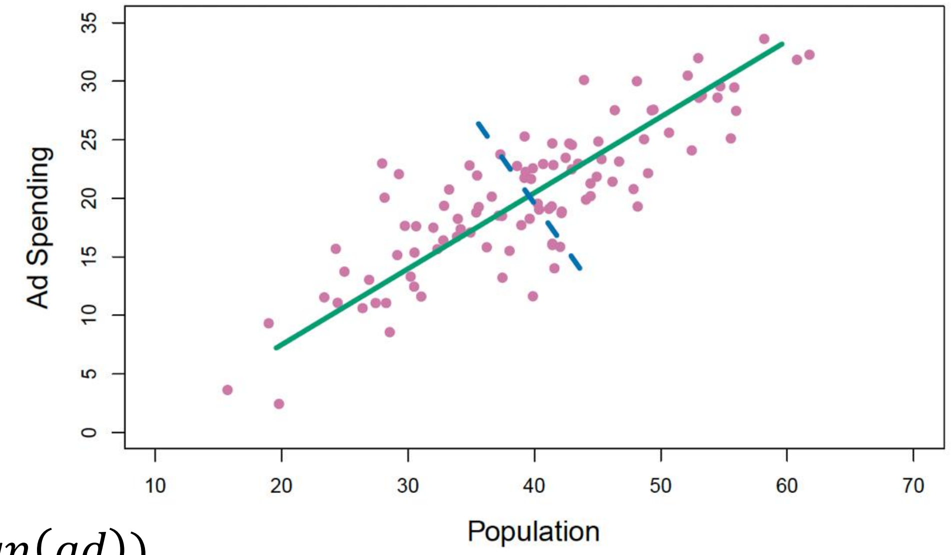
- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA

Let's assume: $\emptyset_{11} = 0.839$, $\emptyset_{21} = 0.544$

Then, $\emptyset_{11} \times \emptyset_{11} + \emptyset_{21} \times \emptyset_{21} = 1$



$$z_i = 0.839 \times (pop_i - mean(pop)) + 0.544 \times (adi - mean(ad))$$

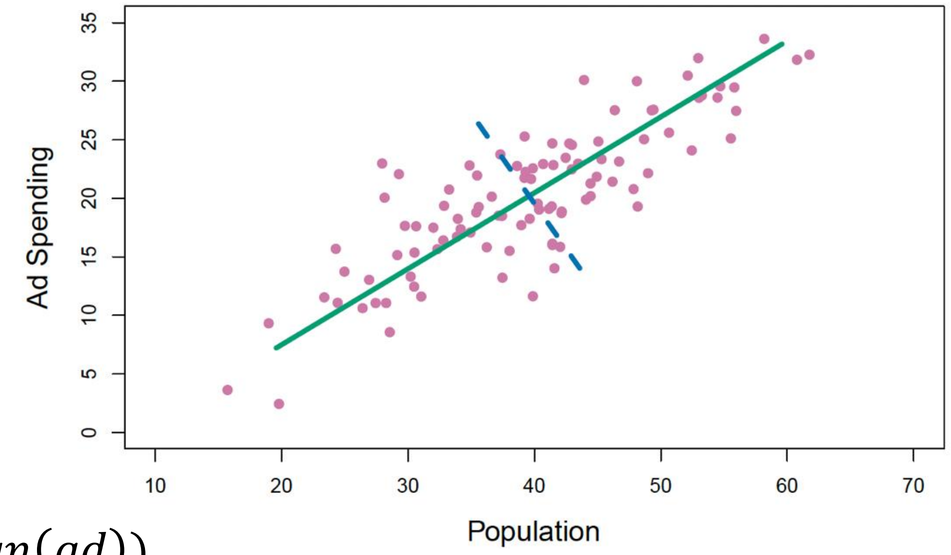
McMaster University

Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11} X_1 + \emptyset_{21} X_2 + \ldots + \emptyset_{p1} X_p$$

  - $\emptyset_{11}$ = loadings of the PCA



Let's assume: $\emptyset_{11} = 0.839$, $\emptyset_{21} = 0.544$

Then, $\emptyset_{11} \times \emptyset_{11} + \emptyset_{21} \times \emptyset_{21} = 1$

Principal component scores

$$z_i = 0.839 \times \left(pop_i - mean(pop)\right) + 0.544 \times (adi - mean(ad))$$

McMaster University

Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA

How do we compute principal components?

Principal Component Analysis

- First Principal Component:
    - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

- $\emptyset_{11}$ = loadings of the PCA
- Normalized linear combination of the features $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

Principal Component Analysis

- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA
  - Normalized linear combination of the features $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
  - Solve the optimization problem:

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

Principal Component Analysis

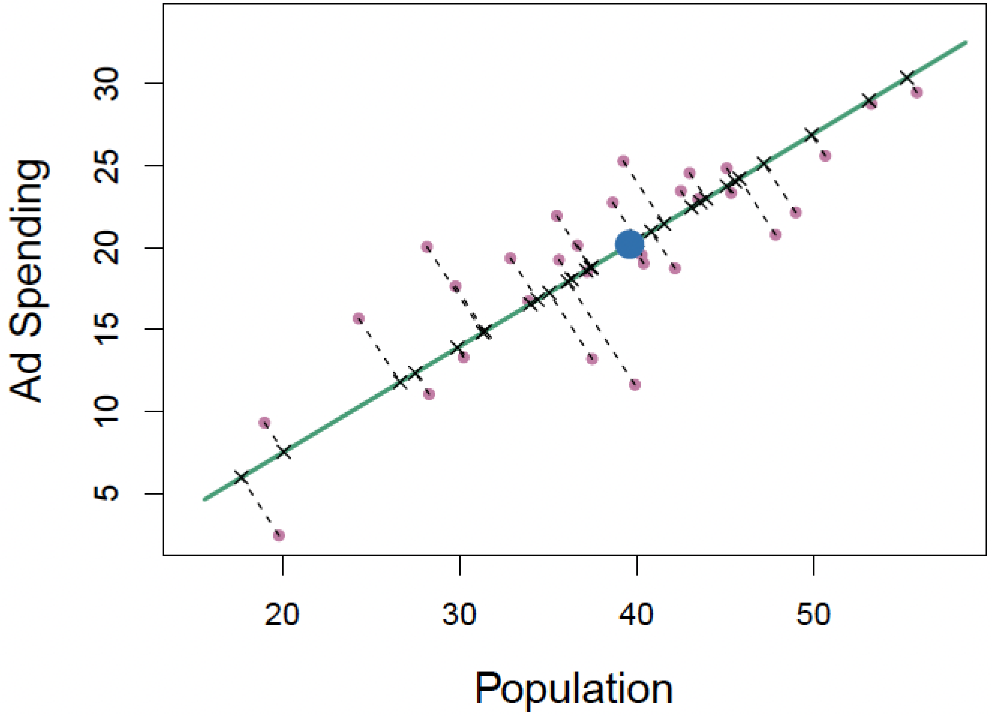
- First Principal Component:
  - Normalized linear combination of the features that has the largest variance.

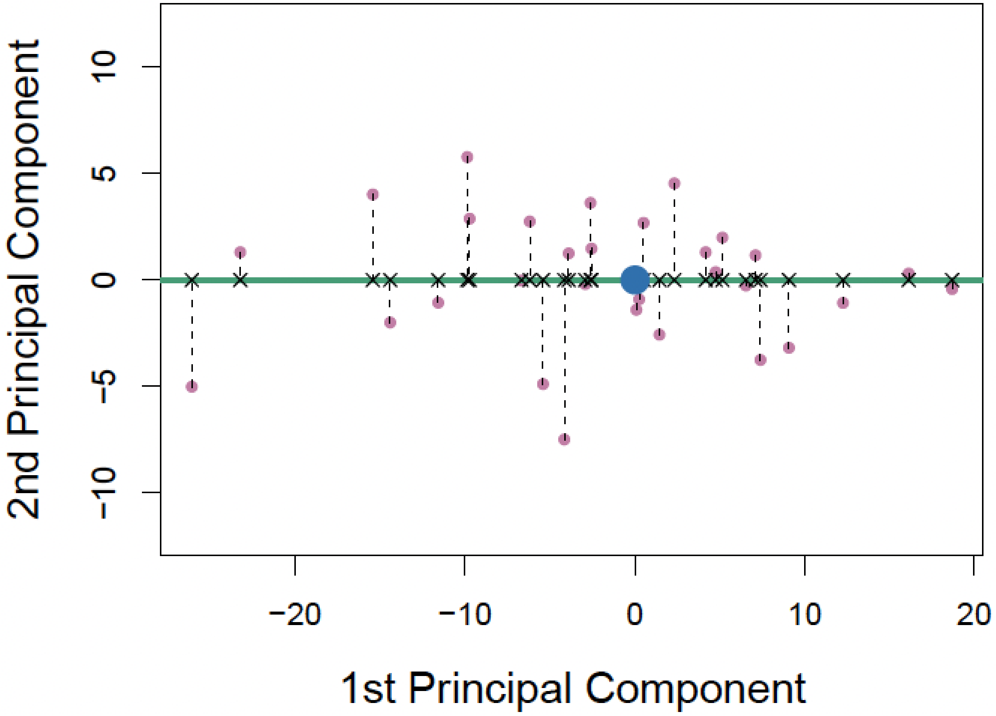
$$Z = \emptyset_{11}X_1 + \emptyset_{21}X_2 + \ldots + \emptyset_{p1}X_p$$

  - $\emptyset_{11}$ = loadings of the PCA
  - Normalized linear combination of the features $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
  - Solve the optimization problem:

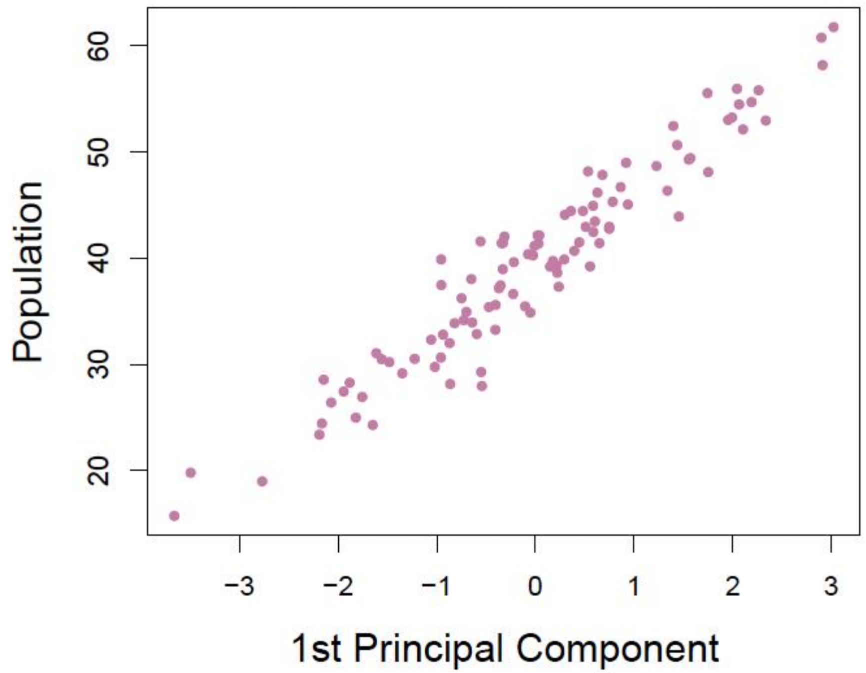
$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \underbrace{\left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2}_{Z_{i1}^2} \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

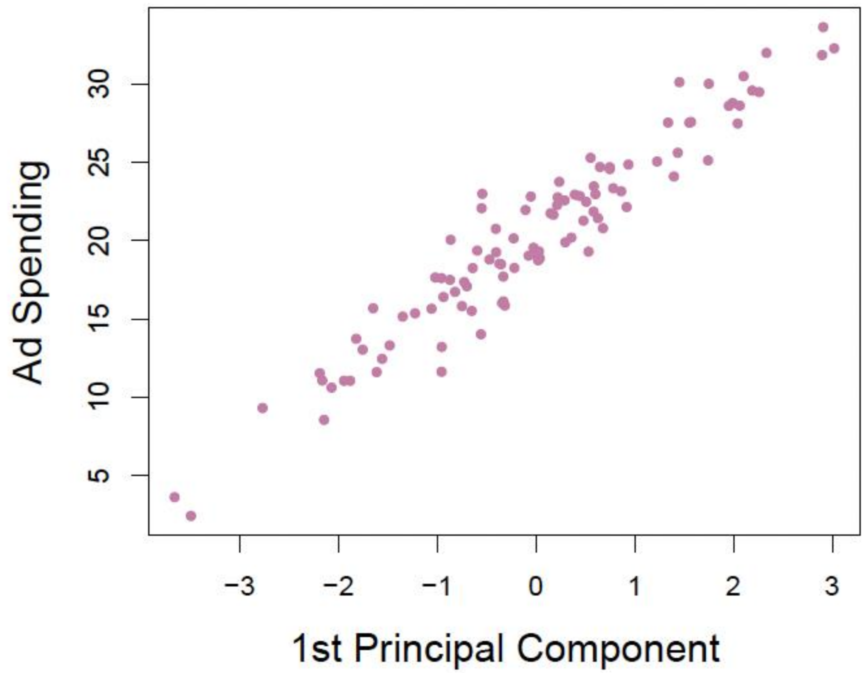
Principal Component Analysis





Principal Component Analysis



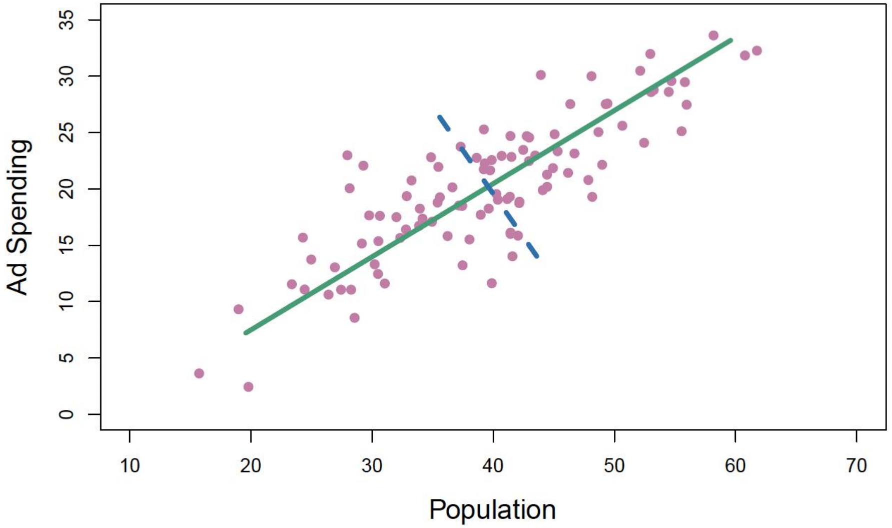


Principal Component Analysis

- First Principal Component (Z_1):
  - Normalized linear combination of the features that has the largest variance.
- Second Principal Component:
  - Normalized linear combination of the features that has the largest variance out of all linear combinations that are uncorrelated with Z_1.

$$Z = \emptyset_{12}X_1 + \emptyset_{22}\, X_2 + \ldots + \emptyset_{p2}X_p$$

  - $\emptyset_{i2}$ = loadings of the PCA



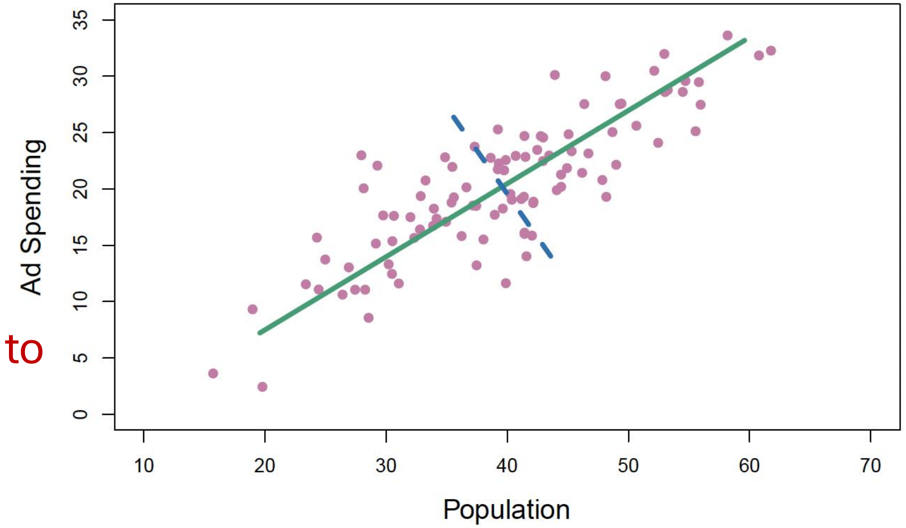
McMaster University

Principal Component Analysis

- First Principal Component (Z_1):
  - Normalized linear combination of the features that has the largest variance.
- Second Principal Component (Z_2)
  - Normalized linear combination of the features that has the largest variance out of all linear combinations that are uncorrelated with Z_1.

$$Z = \emptyset_{12} X_1 + \emptyset_{22} X_2 + \ldots + \emptyset_{p2} X_p$$

  - $\emptyset_{i2}$ = loadings of the PCA
  - Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction $\emptyset_1$ to be orthogonal to the direction of $\emptyset_2$

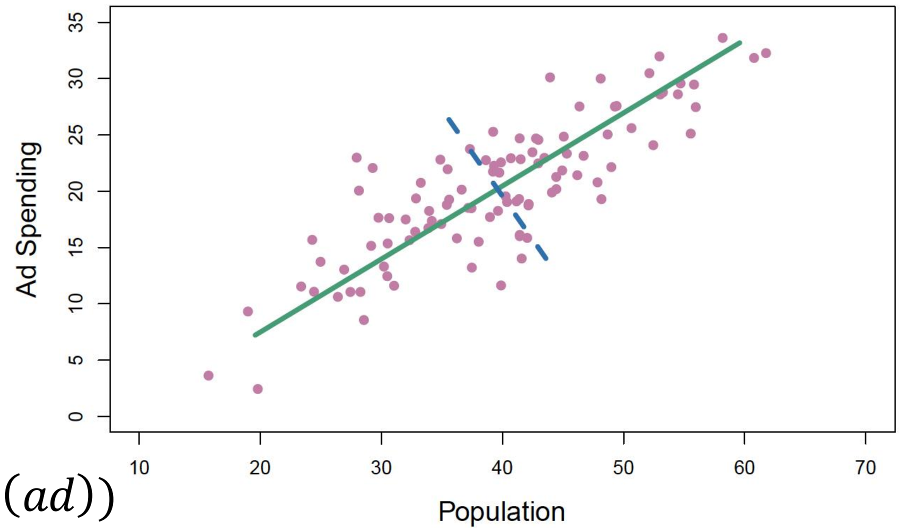


Principal Component Analysis

- First Principal Component (Z_1):
  - Normalized linear combination of the features that has the largest variance.
- Second Principal Component:
  - Normalized linear combination of the features that has the largest variance out of all linear combinations that are uncorrelated with Z_1.

$$Z = \emptyset_{12}X_1 + \emptyset_{22}X_2 + \ldots + \emptyset_{p2}X_p$$

  - $\emptyset_{i2}$ = loadings of the PCA



Second principal component scores

$$z_i = 0.544 \times (pop_i - mean(pop)) - 0.839 \times (ad_i - mean(ad))$$

Principal Component Analysis

- First Principal Component (Z_1):
  - Normalized linear combination of the features that has the largest variance.
- Second Principal Component:
  - Normalized linear combination of the features that has the largest variance out of all linear combinations that are uncorrelated with Z_1.

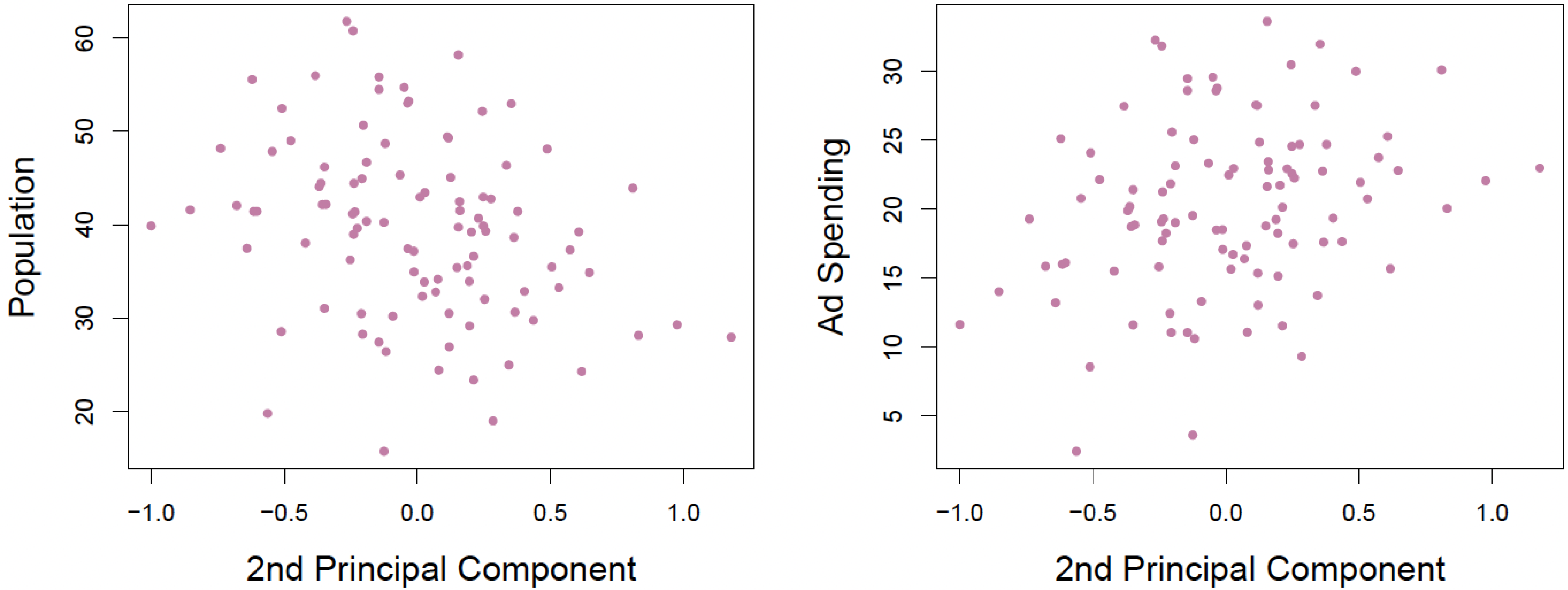
$$Z = \emptyset_{12} X_1 + \emptyset_{22} X_2 + \ldots + \emptyset_{p2} X_p$$

  - $\emptyset_{i2}$ = loadings of the PCA
  - Constraining Z_2 to be uncorrelated with Z_1 is equivalent to constraining the direction $\emptyset_1$ to be orthogonal to the direction of $\emptyset_2$

$$\underset{\phi_{11},\ldots,\phi_{p1}}{\text{maximize}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \boxed{\phi_{j1}} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

Replace with $\emptyset_2$

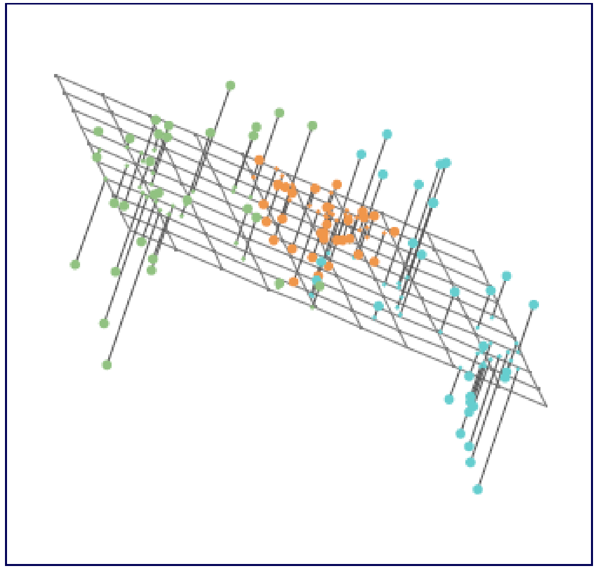
Principal Component Analysis

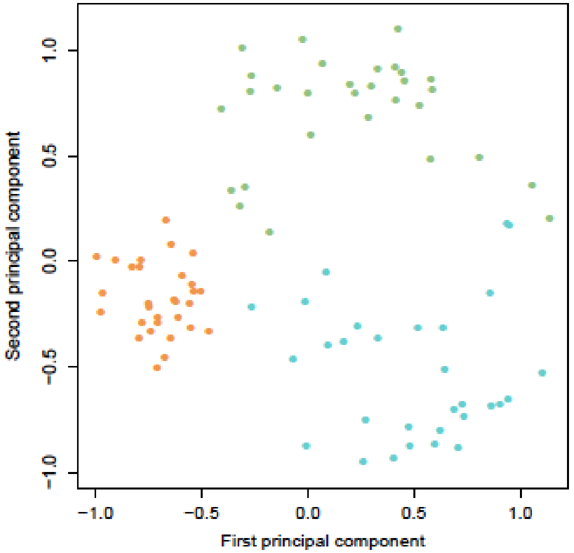


McMaster University

Principal Component Analysis

- For multi-dimensional data, there are multiple principal components.
- Principal components can be used to produce low-dimensional views.

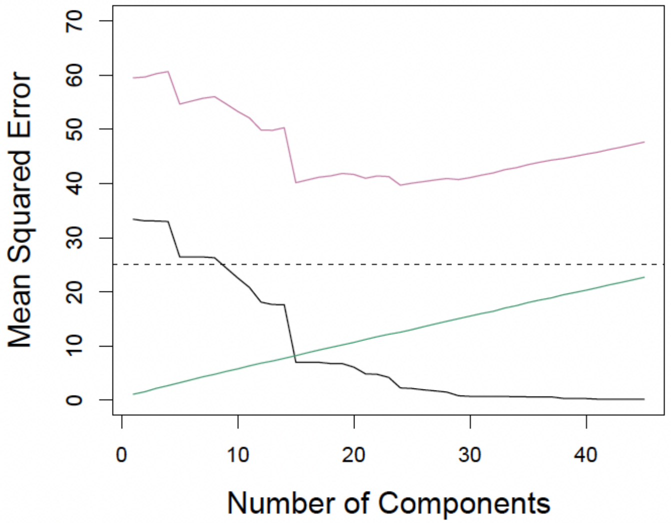


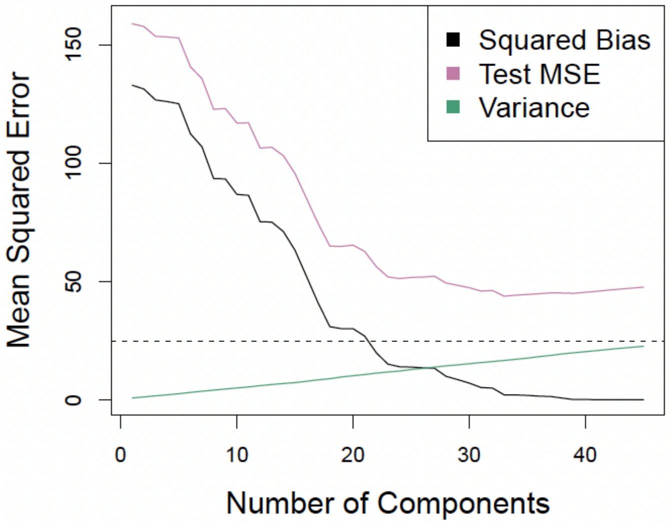


Principal Component Analysis

- For multi-dimensional data, there are multiple principal components.

- Principal components can be used to produce low-dimensional views.
  - It is not feature selection. We consider linear combination of all features





Principal Component Analysis

- For multi-dimensional data, there are multiple principal components.

- Principal components can be used to produce low-dimensional views.
  - It is not feature selection. We consider linear combination of all features

- Eigen decomposition method is used to solve PCA.
  - Principal Component directions are computed using the order sequence of eigenvectors of the matrix $\boldsymbol{X}^T\boldsymbol{X}$

$$\boldsymbol{X}^\top \boldsymbol{X} = \left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top\right)^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top = \boldsymbol{W}\boldsymbol{\Sigma}^2\boldsymbol{W}^\top$$

Principal Component Analysis

- For multi-dimensional data, there are multiple principal components.

- Principal components can be used to produce low-dimensional views.
  - It is not feature selection. We consider linear combination of all features

- Eigen decomposition method is used to solve PCA.
  - Principal Component directions are computed using the order sequence of eigenvectors of the matrix $\boldsymbol{X}^T\boldsymbol{X}$

- For a set of p features and n observations, there are at most $\min(n-1, p)$ principal components. We use the smallest subset, determined by **proportion of variance.**

Principal Component Analysis

- For multi-dimensional data, there are multiple principal components.

- Principal components can be used to produce low-dimensional views.
  - It is not feature selection. We consider linear combination of all features

- Eigen decomposition method is used to solve PCA.
  - Principal Component directions are computed using the order sequence of eigenvectors of the matrix $\boldsymbol{X}^T\boldsymbol{X}$

- For a set of p features and n observations, there are at most $\min(n-1, p)$ principal components. We use the smallest subset, determined by **proportion of variance.**

- Before PCA is performed, the variables should be centered to have mean zero.
  - Scaling heavily impacts PCA result

Principal Component Analysis

- For multi-dimensional data, there are multiple principal components.

- Principal components can be used to produce low-dimensional views.
  - It is not feature selection. We consider linear combination of all features

- Eigen decomposition method is used to solve PCA.
  - Principal Component directions are computed using the order sequence of eigenvectors of the matrix $\boldsymbol{X}^T\boldsymbol{X}$

- For a set of p features and n observations, there are at most $\min(n-1, p)$ principal components. We use the smallest subset, determined by **proportion of variance.**

- Before PCA is performed, the variables should be centered to have mean zero.
  - Scaling heavily impacts PCA result

Popularly used in Recommender systems!

Readings

Required Readings:

Introduction to Statistical Learning

- Chapter 6 – Section 6.3 page 253 – 259
- Chapter 12 – Section 12.2 page 504 – 515

Supplemental Readings (Not required but recommended):

Deep Learning

- Chapter 5 – Section 5.8 page 147 – 150

Thank You

McMaster University