What is Dimension Reduction?

Dimension Reduction is to transform p random variables $X_1, X_2, \ldots, X_p$ to M new random variables $Z_1, Z_2, \ldots, Z_M$ with $M < p$, in the sense of $Z_1, Z_2, \ldots, Z_M$ representing $X_1, X_2, \ldots, X_p$ sufficiently.
This course introduces Linear Dimension Reduction which is Dimension Reduction with each Z_m, $m = 1, 2, \ldots, M$ being linear combination of $X_1, X_2, \ldots, X_p$, i.e.

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j, \tag{1}$$

for some constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$.

Regression Problem: Curse of Dimensionality

Consider the linear regression

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{2}$$

1. As p is large, much flexibility may result in large variance in estimation. Shrinkage Methods can help to reduce flexbility, including the lasso and ridge estimation.
2. Another way is Dimension Reduction, in order to reduce flexibility.

Dimension Reduction on Regression Problem

1. With new random variables $Z_{i1}, Z_{i2}, \ldots, Z_{iM}$, we fit the new regresson model

$$Y_i = \theta_0 + \sum_{m=1}^{M} \theta_m Z_{im} + \eta_i, \quad i = 1, 2, \ldots, n. \tag{3}$$

2. Relation between new coefficients θ_m, $m = 1, 2, \ldots, M$ and original coefficients β_j, $j = 1, 2, \ldots, p$.

$$\begin{aligned} \sum_{m=1}^{M} \theta_m Z_{im} &= \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} X_{ij} \\ = \quad & \sum_{j=1}^{p} \sum_{m=1}^{M} \theta_m \phi_{jm} X_{ij} = \sum_{j=1}^{p} \beta_j X_{ij}. \end{aligned}$$

Hence

$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}, \quad j = 1, 2, \ldots, p. \tag{4}$$

Dimension Reduction Techniques

The pursuit of new variables $Z_1, Z_2, \ldots, Z_M$ plays a key role in Dimension Reduction. Or equivalently,

1. How to find the linear combinations

$$Z_{im} = \sum_{j=1}^{p} \phi_{jm} X_{ij}, \tag{5}$$

   for some constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}$?

2. How to define a reasonal sense under which $Z_{im}, m = 1, 2, \ldots, M$ representing $X_{ij}, j = 1, 2, \ldots, p$?

Two Dimension Reduction Techniques will be introduced.

1. Principal Component Regression (PCR)
2. Partial Least Squares (PLS)

Principal Component Regression (PCR)

PCR is a dimension reduction technique to solve regression problem. It has two steps.

1. The key idea is to find the new predictors $Z_{im}, m = 1, 2, \ldots, M$ that can explain most of the variability in the data. This procedure is based on Principal Component Analysis (PCA).
2. Then fit the new regression model, in which $Z_{im}, m = 1, 2, \ldots, M$ are predictors. The estimation for original coefficients $\beta_j, j = 1, 2, \ldots, p$ can be obtained via (4) and estimation for new coefficients $\theta_m, m = 1, 2, \ldots, M$.

Principal Component Analysis (PCA)

Consider p correlated random variables $X_1, X_2, \ldots, X_p$ with observations $X_{i1}, X_{i2}, \ldots, X_{ip}$, where $i = 1, 2, \ldots, n$. For convenience, assume $\mathbb{E}(X_j) = 0$.

1. PCA is a technique to summarize the p random variables with a smaller number of representative variables that collectively explain most of the variability in the original set.
2. PCA is an unsupervised learning method while PCR is supervised learning. PCR applies PCA on regression problems.

Procedure of PCA (1)

PCA provides a set of new variables $Z_{i1}, Z_{i2}, \ldots, Z_{iq}$ with $q = \min(p, n)$ in the following way.

(1). $Z_{i1} = \sum_{j=1}^{p} \phi_{1j} X_{ij}$ where $\phi_{1j}, j = 1, 2, \ldots, p$ is the maximizer of the optimization problem

$$\max_{\phi_{11}, \phi_{12}, \ldots, \phi_{1p}} \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{1j} X_{ij} \right)^2$$

$$s.t. \sum_{j=1}^{p} \phi_{1j}^2 = 1 \tag{6}$$

Procedure of PCA (2)

Denote $\boldsymbol{\phi}_k = (\phi_{k1}, \phi_{k2}, \ldots, \phi_{kp})$ with $k = 1, 2, \ldots, q$ and $q = \min(p, n)$.

(2). $Z_{i2} = \sum_{j=1}^{p} \phi_{2j} X_{ij}$ where $\phi_{2j}, j = 1, 2, \ldots, p$ is the maximizer of the optimization problem (6) with additional constraint $\boldsymbol{\phi}_2$ is orthogonal to $\boldsymbol{\phi}_1$.

(3). $Z_{i3} = \sum_{j=1}^{p} \phi_{3j} X_{ij}$ where $\phi_{3j}, j = 1, 2, \ldots, p$ is the maximizer of the optimization problem (6) with additional constraint $\boldsymbol{\phi}_3$ is orthogonal to $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$.

(4). Continue this procedure until construction of Z_{iq} with $q = \min(p, n)$.

Elements of PCA

Two common used concepts in PCA as follows.

1. Principal Component Scores: $Z_{i1}, Z_{i2}, \ldots, Z_{iq}$.
2. Principal Component Loading Vectors: $\phi_1, \phi_2, \ldots, \phi_q$.

Remark

The elements of each principal component loading vector are called principal component loadings. Usually, principal component scores are also named as principal components for simplicity.

Representation by Principal Components

Fortunately, Mercer' Theorem shows that original data X_{ij} can be represented by principal component scores and principal component loading vectors in the following way.

$$X_{ij} = \sum_{m=1}^{q} Z_{im}\phi_{mj}, \quad i = 1, 2, \ldots, n; \; j = 1, 2, \ldots, p. \tag{7}$$

Note that $q = \min(p, n)$.
This representation illustrates that principal component scores contain sufficient information in the original data.

Approximated Representation

Our initial aim is to pursue small number of new variables to represent original data.
Fortunately again, the first M principal components scores and loading vectors can represent original data approximately

$$X_{ij} \approx \sum_{m=1}^{M} Z_{im}\phi_{mj}, \quad i = 1, 2, \ldots, n; \; j = 1, 2, \ldots, p, \tag{8}$$

in the sense of keeping most variability given an integer M.

Example 1: Principal Component Loading Vector

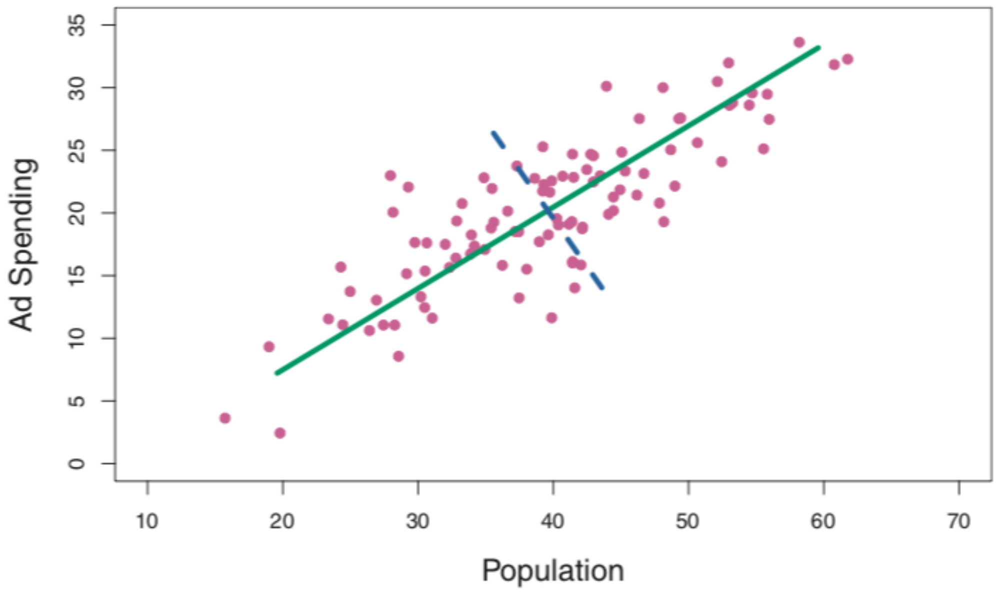


FIGURE 6.14. *The population size (`pop`) and ad spending (`ad`) for 100 different cities are shown as purple circles. The green solid line indicates the first principal component, and the blue dashed line indicates the second principal component.*

Example 1: Principal Component Scores

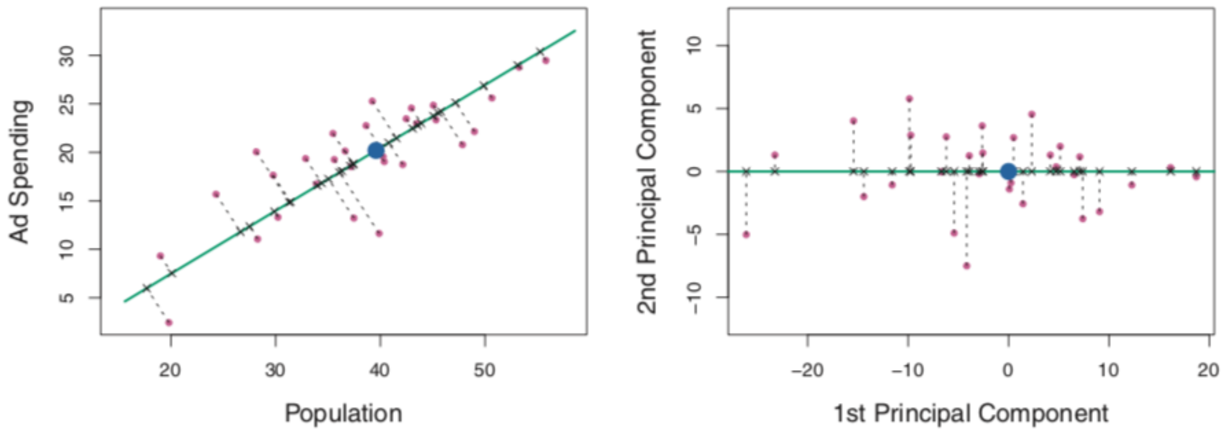


FIGURE 6.15. *A subset of the advertising data. The mean* pop *and* ad *budgets are indicated with a blue circle.* Left: *The first principal component direction is shown in green. It is the dimension along which the data vary the most, and it also defines the line that is closest to all n of the observations. The distances from each observation to the principal component are represented using the black dashed line segments. The blue dot represents* $(\overline{\texttt{pop}}, \overline{\texttt{ad}})$. Right: *The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.*

Example 1: First Principal Component Score

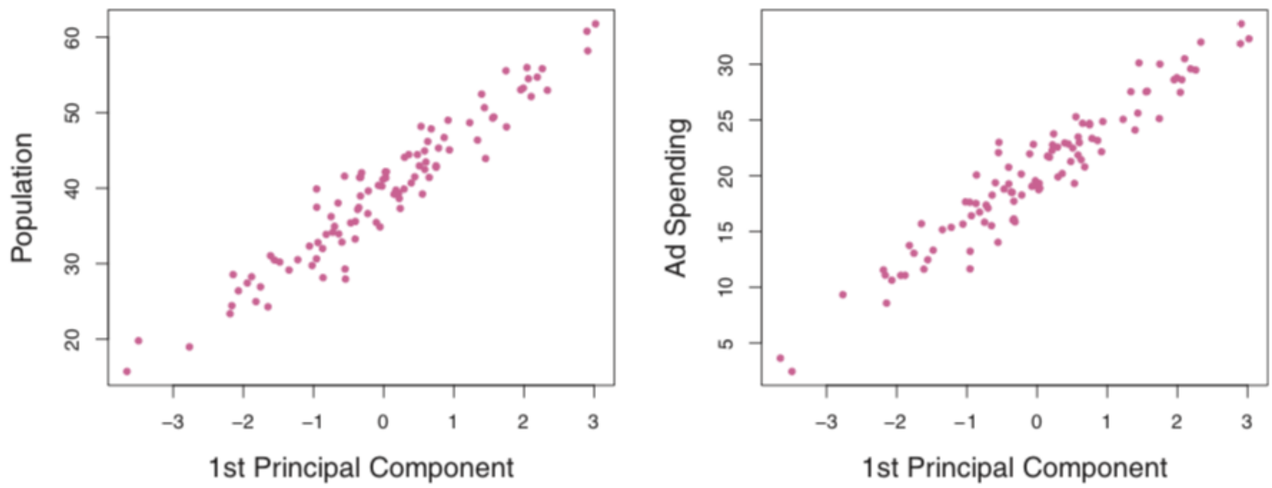


FIGURE 6.16. *Plots of the first principal component scores z_{i1} versus* pop *and* ad*. The relationships are strong.*

Example 1: Second Principal Component Score

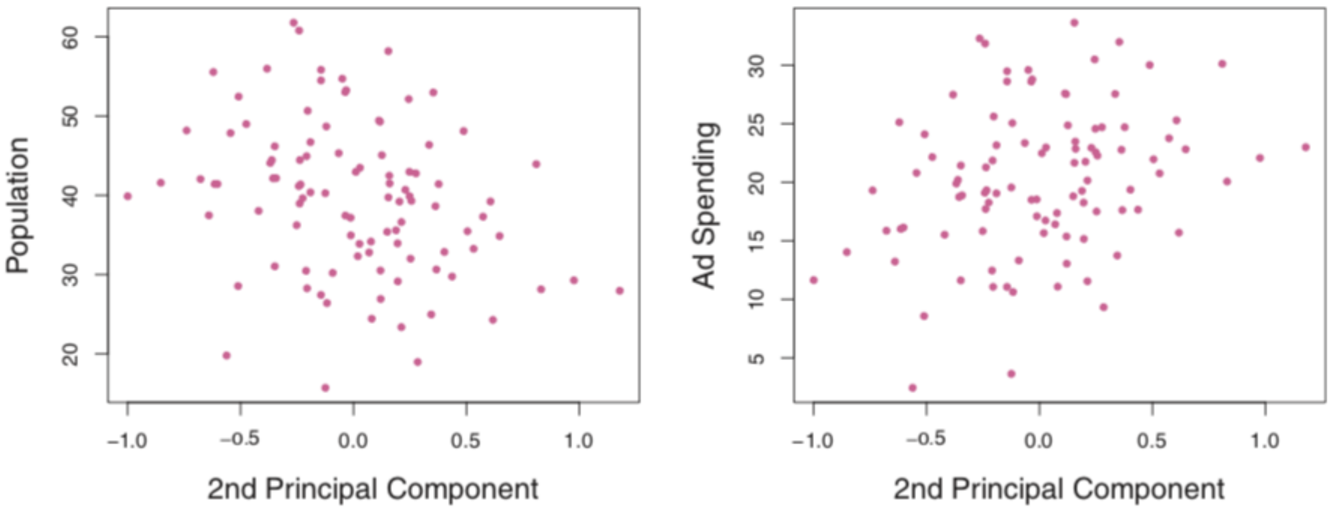


FIGURE 6.17. *Plots of the second principal component scores* z_{i2} *versus* `pop` *and* `ad`. *The relationships are weak.*

PCA: Interpretation

Two ways to interprete PCA

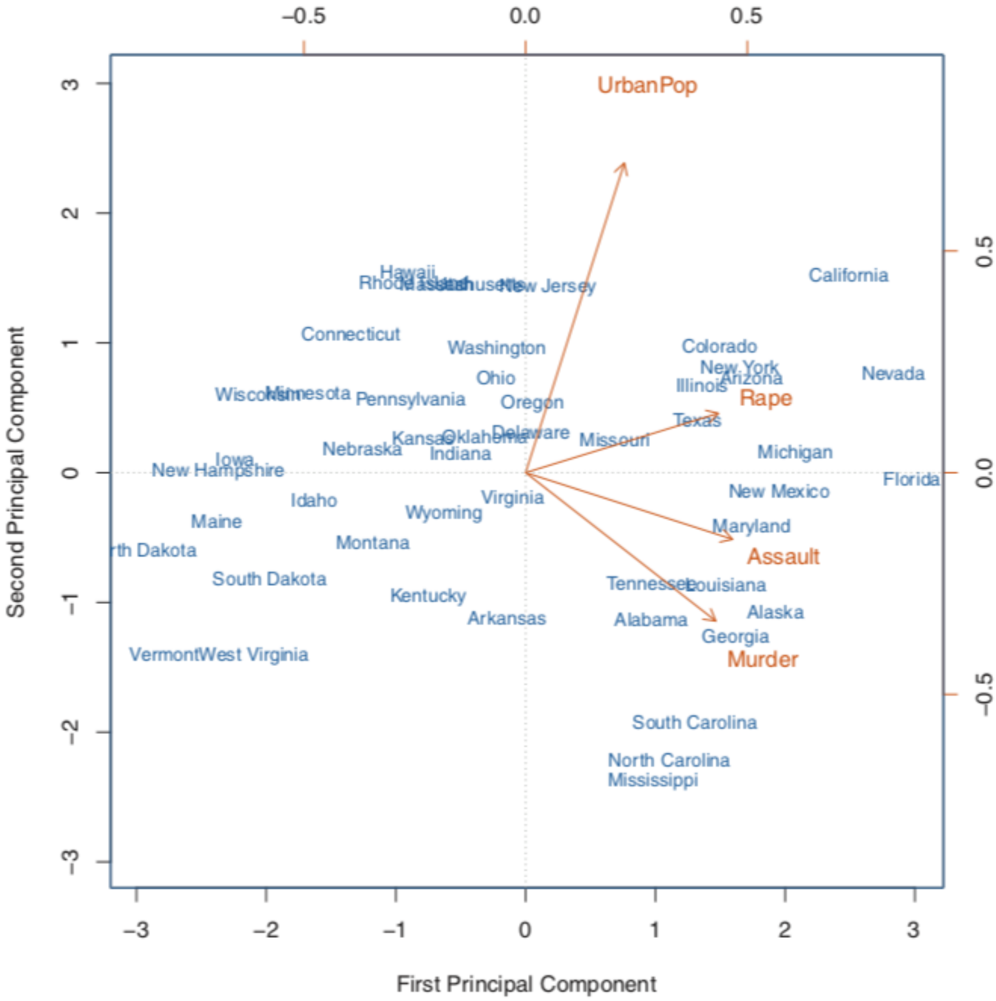
1. Principal component loading vectors are the directions in feature space along which the data vary the most, and the principal component scores are projections along these directions.
2. PCA provides low-dimensional linear surfaces that are closest to the observations.

Example 2: USArrests Data

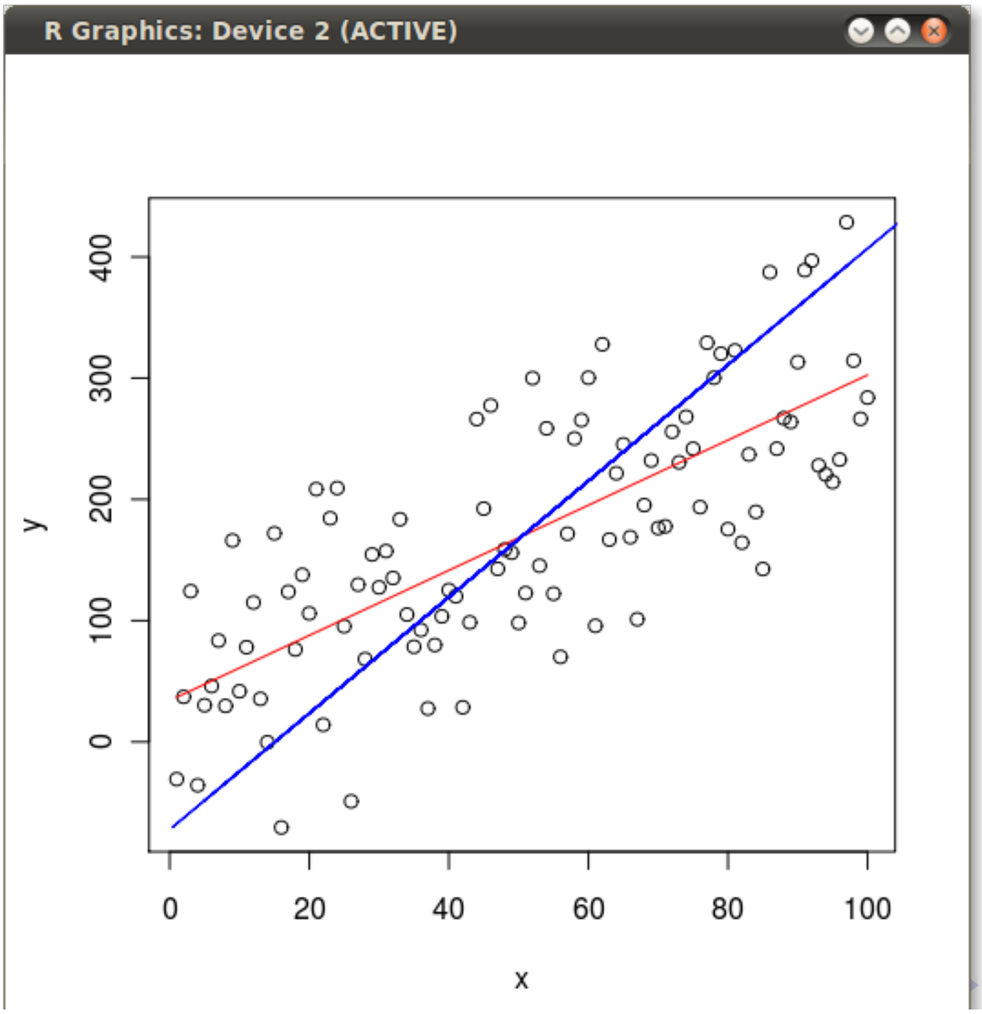
	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

TABLE 10.1. *The principal component loading vectors, ϕ_1 and ϕ_2, for the* `USArrests` *data. These are also displayed in Figure 10.1.*

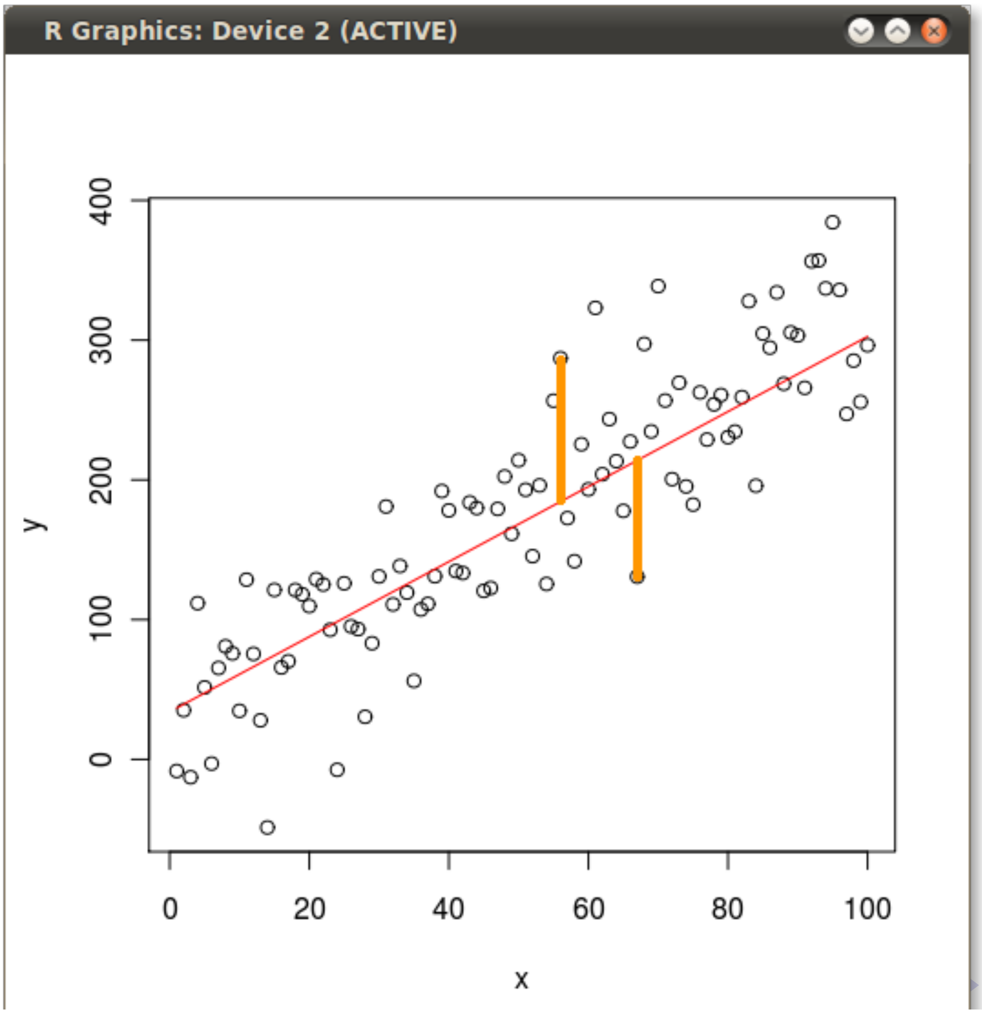
Example 2: A Biplot



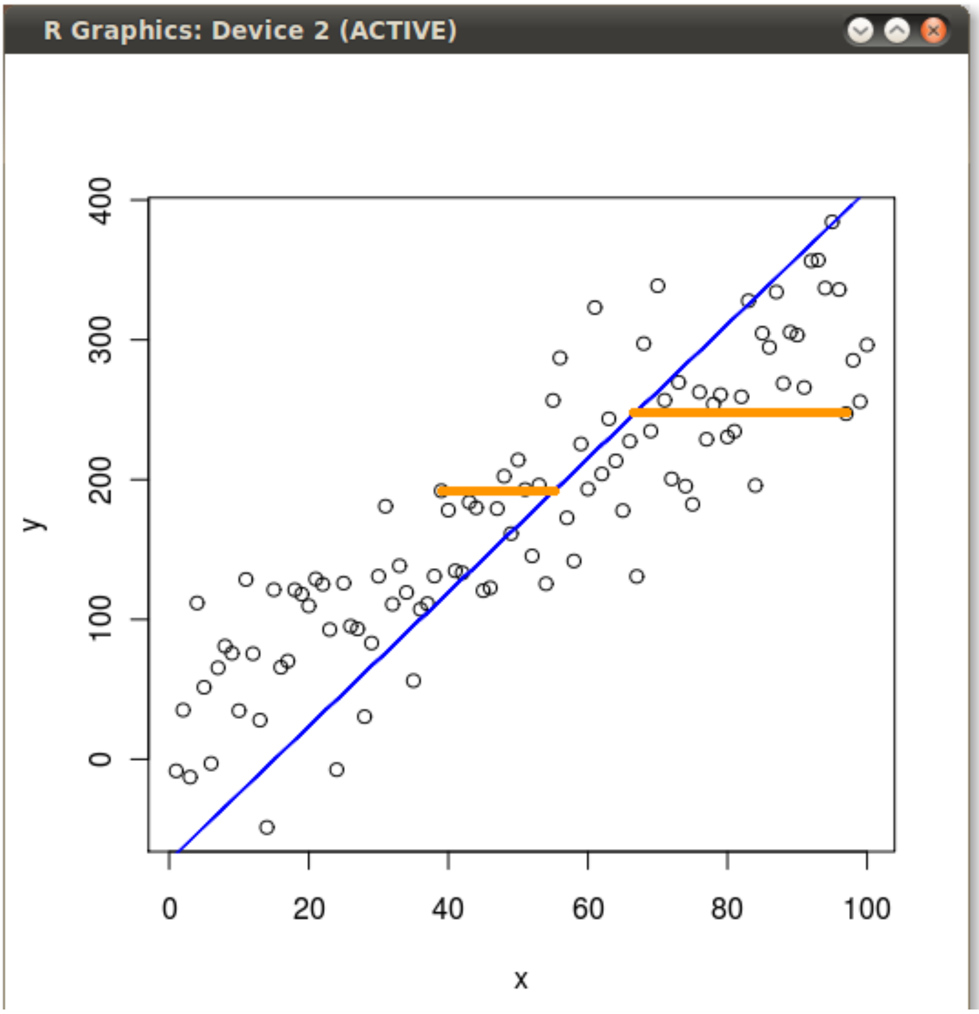
OLS: $y \sim x$ and $x \sim y$



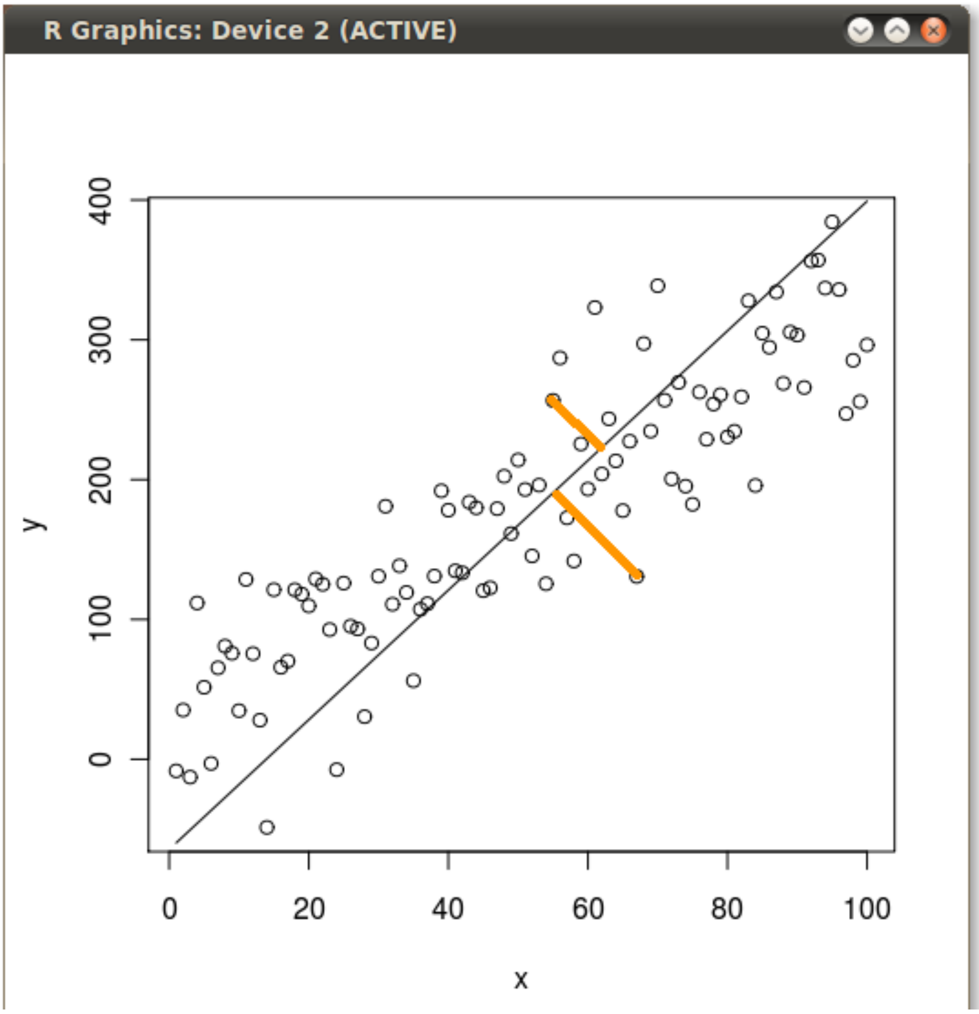
Geometry of OLS: $y \sim x$



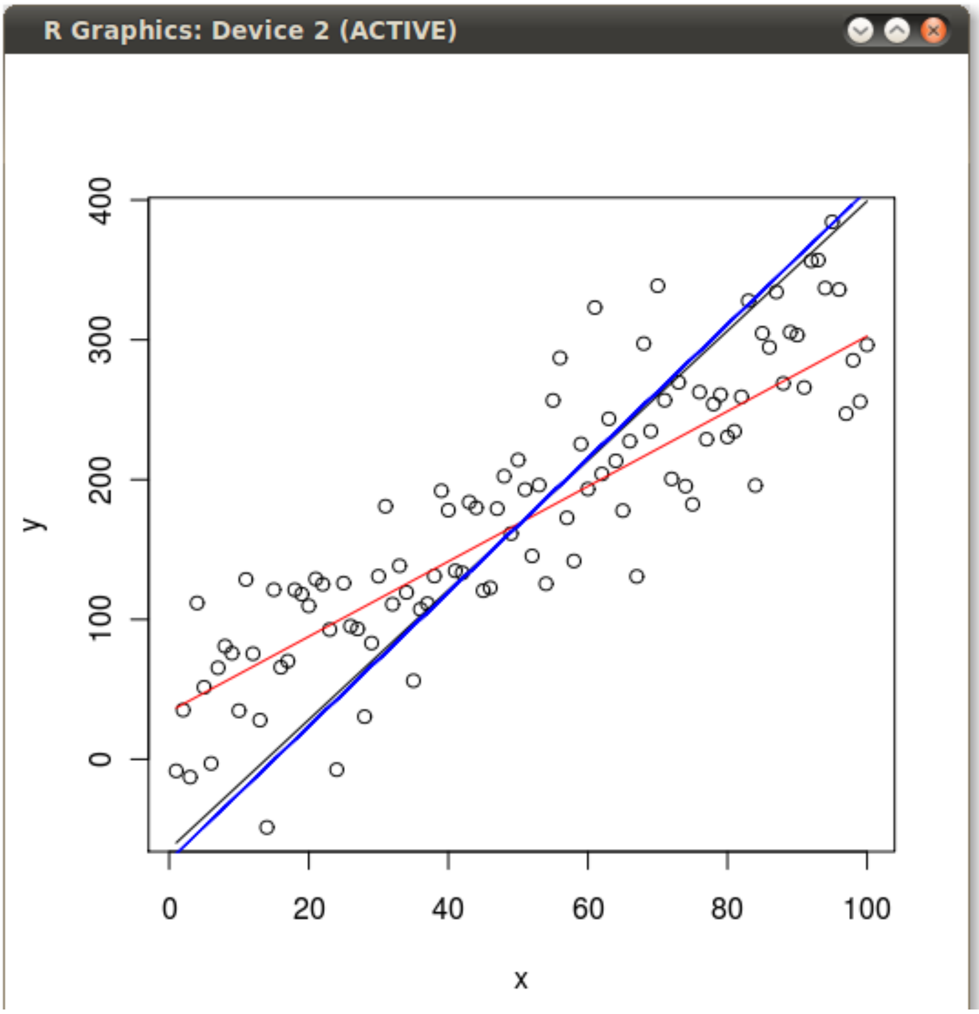
Geometry of OLS: $x \sim y$



Geometry of PCA



Comparison Between PCA and OLS



Example 1: Drawback of PCR in Regression

Consider a linear regression model

$$y_i = 0.2x_{1i} + 0.4x_{2i} + 1.6x_{3i} + \varepsilon_i, \quad i = 1, 2, \ldots, 100, \tag{1}$$

where x_{1i}, x_{2i}, x_{3i} are from normal distribution with zero mean and unit variance. The correlation between x_{1i} and x_{2i} is 0.9. x_{3i} is independent of x_{1i} and x_{2i}.

Analysis on PCA Results:

The covariance matrix of x_{1i}, x_{2i}, x_{3i} is

$$\begin{pmatrix} 1 & 0.9 & 0 \\ 0.9 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

The eigenvalues of this matrix are $1.9, 1, 0.1$ and the corresponding eigenvectors are

$$\begin{pmatrix} 0.707 \\ 0.707 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -0.707 \\ 0.707 \\ 0 \end{pmatrix} \tag{3}$$

PLS

Partial Least Squares (PLS) is a supervised alternative to PCR. For a linear regression model

$$Y_i = X_{1i}\beta_1 + X_{2i}\beta_2 + \ldots + X_{pi}\beta_p, \quad i = 1, 2, \ldots, n. \tag{4}$$

PLS will pursue a set of new predictors Z_{mi}, $i = 1, 2, \ldots, n$ and then regress Y_i on this set of new predictors.

1. Simple linear regression by regressing Y_i on each predictor X_{ji} and then get ϕ_{1j}. The first new predictor is $Z_{1i} = \sum_{j=1}^{p} \phi_{1j} X_{ji}$.
2. Remove Information about Z_{1i} from Original Data. Regression Each predictor X_{ji} on Z_{1i} and take the residuals as updated predictors. Then repeat the first step on the updated predictors. Then Z_{2i} is obtained.
3. Repeat Step 2 and get a set of new predictors $Z_{1i}, Z_{2i}, \ldots, Z_{Mi}$.