Dimension Reduction Methods

- The methods that we have discussed so far in this chapter have controlled variance in two different ways, either by using a subset of the original variables, or by shrinking their coefficients toward zero. All of these methods are defined using the original predictors, $X_1, X_2, \ldots, X_p$.
- We now explore a class of approaches that transform the predictors and then fit a least squares model using the transformed variables. We will refer to these techniques as *dimension reduction* methods.
- Let $Z_1, Z_2, \ldots, Z_M$ represent $M < p$ linear combinations of our original p predictors. That is,

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j \longrightarrow z_{im} = \sum_{j=1}^{p} \phi_{jm} x_{ij}$$

  for some constants $\phi_{1m}, \phi_{2m}, \ldots, \phi_{pm}, m = 1, ..., M$.
- We can then fit the linear regression model

$$y_i = \theta_0 + \sum_{m=1}^{M} \theta_m z_{im} + \epsilon_i, \quad i = 1, \ldots, n,$$

  using least squares.

$\theta_0, \theta_1, \cdots, \theta_M$: regression coefficients

- If the constants $\phi_{1m}, \phi_{2m}, ..., \phi_{pm}$ are chosen wisely, then such dimension reduction approaches can often outperform least squares regression.

In other words, fitting

$$Y = \theta_0 + \theta_1 Z_1 + \theta_2 Z_2 + \cdots + \theta_M Z_M + \varepsilon$$

using least squares can lead to better results than fitting

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

using least squares.

- The term dimension reduction comes from the fact that this approach reduces the problem of estimating the $p+1$ coefficients $\beta_0, \beta_1, ..., \beta_p$ to the simpler problem of estimating the $M+1$ coefficients $\theta_0, \theta_1, ..., \theta_M$, where $M < p$. In other words, the dimension of the problem has been reduced from $p+1$ to $M+1$.

note that

$$\sum_{m=1}^{M} \theta_m z_{im} = \sum_{m=1}^{M} \theta_m \sum_{j=1}^{p} \phi_{jm} x_{ij} = \sum_{j=1}^{p} \left(\sum_{m=1}^{M} \theta_m \phi_{jm} \right) x_{ij}$$

$$= \sum_{j=1}^{p} \beta_j x_{ij}$$

where

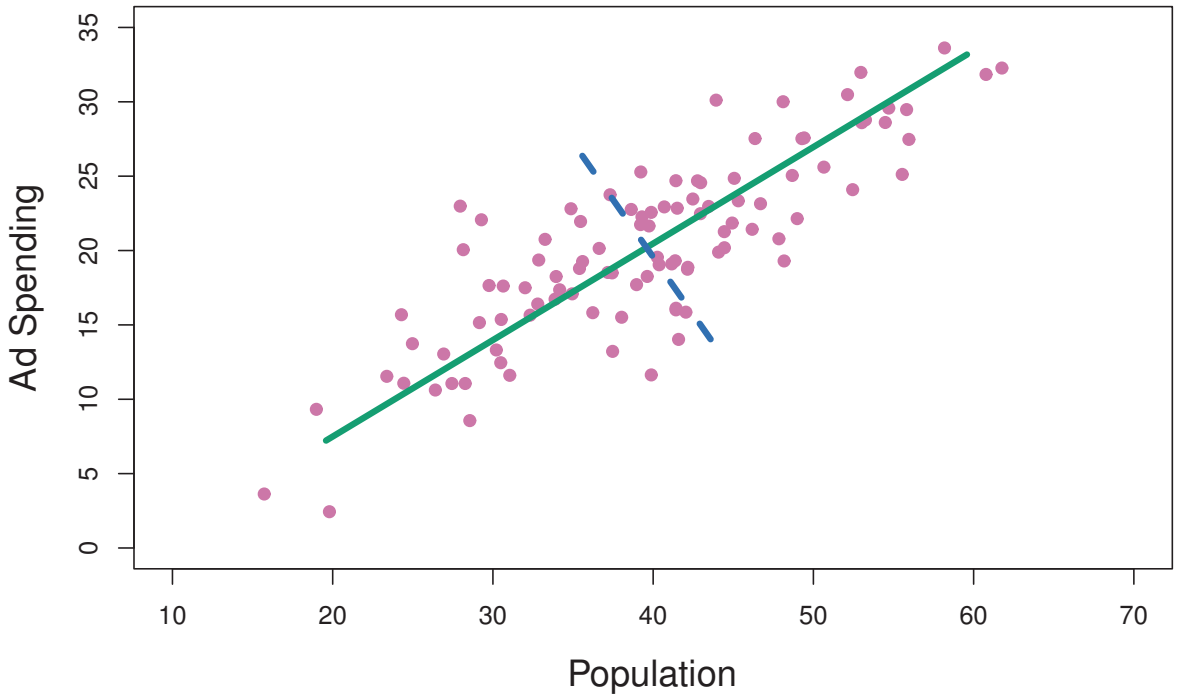
$$\beta_j = \sum_{m=1}^{M} \theta_m \phi_{jm}$$

note: when p is large relative to n, selecting a value of $M \ll p$ can significantly reduce the variance of the fitted coefficients.

- All dimension reduction methods work in two steps:
  - First, the transformed predictors $Z_1, Z_2, ..., Z_M$ are obtained.
  - Second, the model is fit using these M predictors.
- However, the choice of $Z_1, Z_2, ..., Z_M$, or equivalently, the selection of the ϕ_{jm}'s, can be achieved in different ways. Here, we will consider two approaches for this task: *principal components* and *partial least squares*.

An Overview of Principal Components Analysis

- Principal components analysis (PCA) is a popular approach for deriving a low-dimensional set of features from a large set of variables. Here we describe its use as a dimension reduction technique for regression.
- The first principal component is that (normalized) linear combination of the variables with the largest variance.
- The second principal component has largest variance, subject to being uncorrelated with the first.
- And so on.
- Hence with many correlated original variables, we replace them with a small set of principal components that capture their joint variation.



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.

The first principal component is given by

$$Z_1 = 0.839 \times (pop - \overline{pop}) + 0.544 \times (ad - \overline{ad})$$

(ϕ_{11} = 0.839, ϕ_{21} = 0.544)

ϕ_{11} and ϕ_{21} are the principal component loadings

$$Z_1 = \begin{bmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{bmatrix}, \quad n = 100,$$

$\overline{pop}$: mean of all pop values

$\overline{ad}$: mean of all ad values

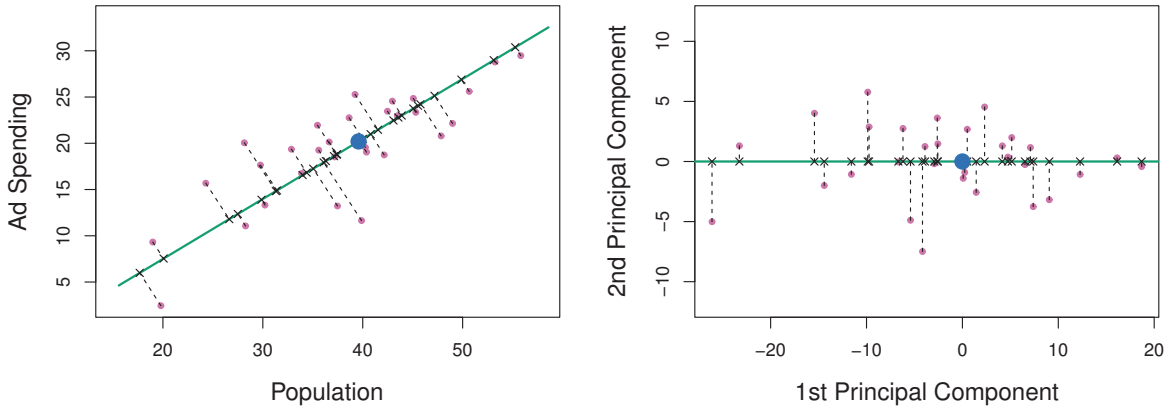
$$z_{i1} = 0.839 \times (pop_i - \overline{pop}) + 0.544 \times (ad_i - \overline{ad})$$

$z_{11}, z_{21}, \ldots, z_{n1}$ are known as the principal component scores.

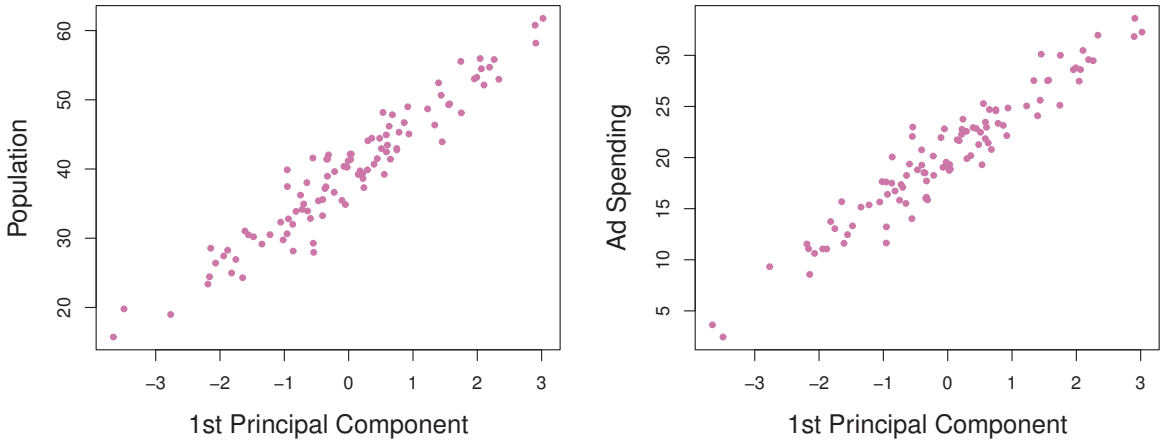
ϕ_{11} and ϕ_{21} are the ones that maximize

$$\text{Var}\left(\phi_{11} \times (\text{pop} - \overline{\text{pop}}) + \phi_{21} \times (\text{ad} - \overline{\text{ad}})\right)$$

$$\text{subject to } \phi_{11}^2 + \phi_{21}^2 = 1$$



A subset of the advertising data. 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 $(\overline{\text{pop}}, \overline{\text{ad}})$. Right: The left-hand panel has been rotated so that the first principal component direction coincides with the x-axis.



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

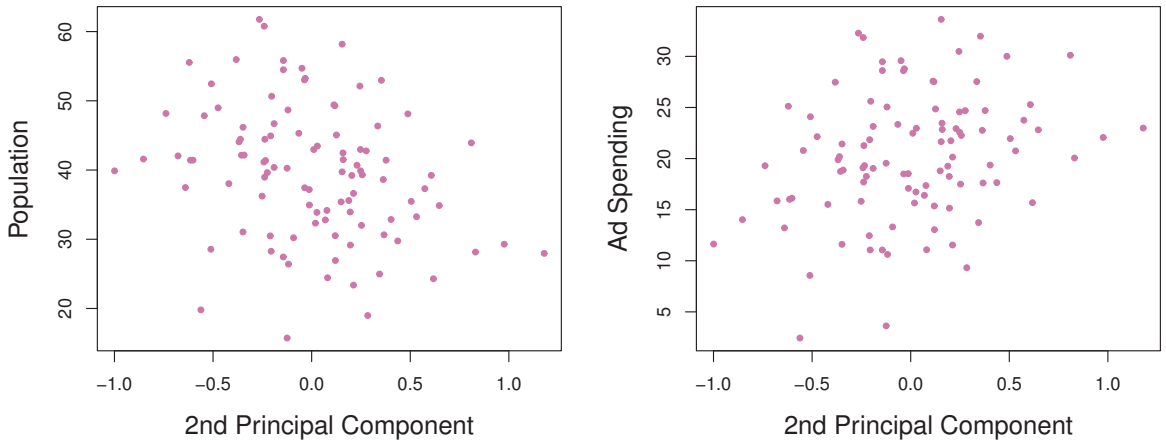
The second principal component Z_2 is a linear combination of the variables that is uncorrelated with Z_1, and has largest variance.

In our example:

$$Z_2 = 0.544 \times (pop - \overline{pop}) - 0.839 \times (ad - \overline{ad})$$

The zero correlation condition of Z_1 and Z_2 is equivalent to the condition that the direction must be perpendicular, or orthogonal, to the first principal component direction.

In general, one can construct up to p distinct principal components.



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

The Principal Components Regression Approach

- The principal components regression (PCR) approach involves constructing the first M principal components, $Z_1, ..., Z_M$, and then using these components as the predictors in a linear regression model that is fit using least squares.
- The key idea is that often a small number of principal components suffice to explain most of the variability in the data, as well as the relationship with the response.
- In other words, we assume that the directions in which $X_1, \ldots, X_p$ show the most variation are the directions that are associated with Y.
- If the assumption underlying PCR holds, then fitting a least squares model to $Z_1, ..., Z_M$ will lead to better results than fitting a least squares model to $X_1, ..., X_p$,
- since most or all of the information in the data that relates to the response is contained in $Z_1, ..., Z_M$, and by estimating only $M \ll p$ coefficients we can mitigate overfitting.