# Principal Components Analysis (PCA)

# Principal Components Analysis (PCA)

- In the previous plot, the **First PC** is actually:

$$M1 = 0.839\ (pop\ -\ \overline{pop}) + 0.544\ (ad\ -\ \overline{ad})$$

- Notice that the **PCs** are **centered** on the **means** of pop and ad
- **0.839** and **0.544** called the **"Loadings"** of Principal Component 1, and they are the weights of the pop and ad variables, respectively, which establish the direction of **$M_1$**
- Notice an important **property** → **$0.839^2 + 0.544^2 = 1$**
- More generally:
  - For **P variables** $x_1, x_2, \ldots x_P$, there are **P principal components** $M_1, M_2, \ldots M_P$
  - **Each** principal components has **P loadings**
  - The loadings are **scaled** so that the **sum** of the **squared loadings** for each component **= 1**
  - All components are **"orthogonal"** (i.e., perpendicular, independent or uncorrelated) with each other

# Variables, Components & Scores

- For a set of **P variables** → $X_1, X_2, \ldots\ldots X_P$
- There are **P "orthogonal" PC's**:

$$M_1 = l_{11}X_1 + l_{12}X_2 + \cdots + l_{1P}X_P$$
$$M_2 = l_{21}X_1 + l_{22}X_2 + \cdots + l_{2P}X_P$$
$$\ldots\ldots$$
$$M_P = l_{P1}X_1 + l_{P2}X_2 + \cdots + l_{PP}X_P$$

- Each with a **sum** of **squared** factor **loadings** = 1:

$$l_{11}^2 + l_{12}^2 + \cdots + l_{1P}^2 = 1$$
$$l_{21}^2 + l_{22}^2 + \cdots + l_{2P}^2 = 1$$
$$\ldots\ldots$$
$$l_{P1}^2 + l_{P2}^2 + \cdots + l_{PP}^2 = 1$$

- For a data point $X_{i1}, X_{i2}, \ldots\ldots X_{ip}$ the corresponding value $m_{i1}$ is called the **PC "score"** for that data point

# Principal Components Regression (PCR)

- The main purpose of **PCR** is to **reduce** the **dimensionality** (i.e., number of variables) of a model **without removing variables**.
- Since the PC's are sorted from highest to lowest variance, it is very likely that the **first few PC's** are **sufficient** to represent the variance in the data.
- So, for a given **OLS** regression:

$$Y = \beta_0 + \beta_1(X_1) + \beta_2(X_2) + \ldots\ldots + \beta_P(X_P) + \varepsilon$$

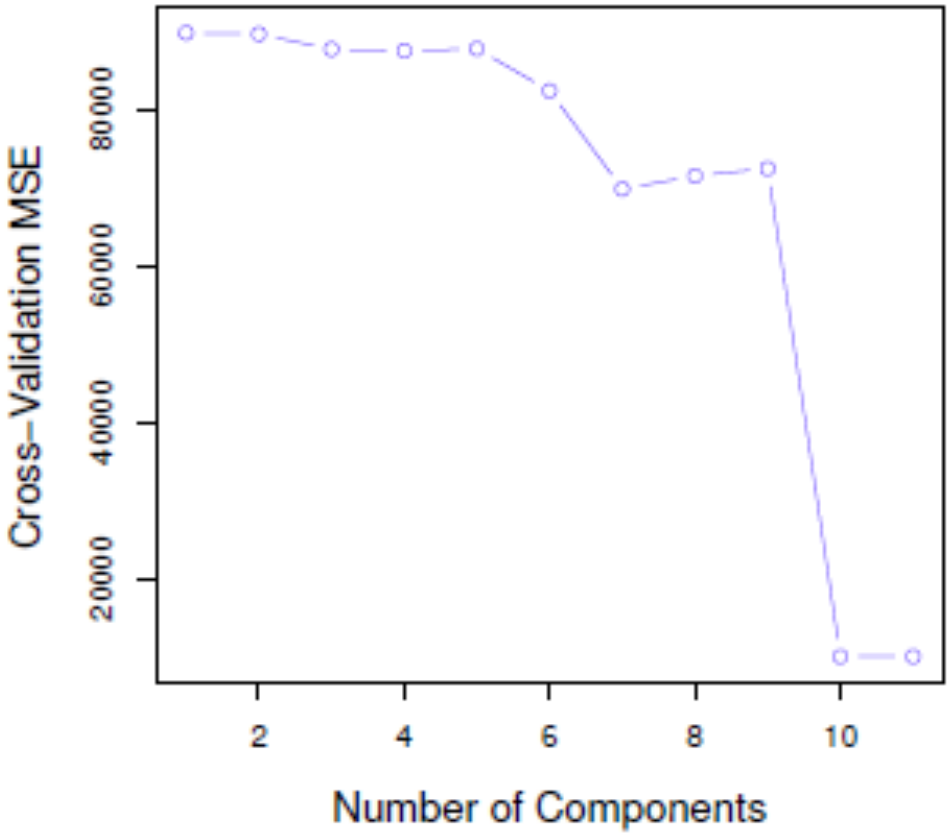
- We can construct **another OLS** model on M PC's:

$$Y = \alpha_0 + \alpha_1(M_1) + \alpha_2(M_2) + \ldots\ldots + \alpha_M(M_M) + \varepsilon$$

- **M** can be anywhere between **1** and **P** (all PC's → same as OLS)
- Since each PC is a linear transformation of all variables using factor loadings, **all variables** are **represented**
- The key is to find the **optimal** number of **M PC's** (i.e., a **"tuning"** parameter) to include in the model
- As with regular OLS, as **M** ↑ → **Bias** ↓ and **Variance** ↑

# Choosing the Number of Directions M

- As with other methods **cross-validation** is the best selector for **M**
- In the example below using the Credit data set, we see that the **10 K-Fold MSE** drops sharply from 9 to 10 PC's and then stays flat.



# Tips

`pcr(){pls}` → The `pcr()` function in the `{pls}` package estimates **PCR** regression models

```
pcr.fit=pcr(y~x1+x2+x3+etc.,data=dataName,
            scale=TRUE,validation="CV")
```
→ `scale=TRUE` standardizes predictors, which is necessary when variables are in different scales (e.g., lbs, feet, etc.); `validation="CV"` does 10-fold cross validation; `validation="LOO"` does leave-one-out cross validation

`validationplot(pcr.fit,val.type="MSEP")` → Print Scree Plot using MSE's

KOGOD SCHOOL
*of*
BUSINESS