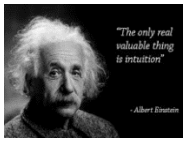
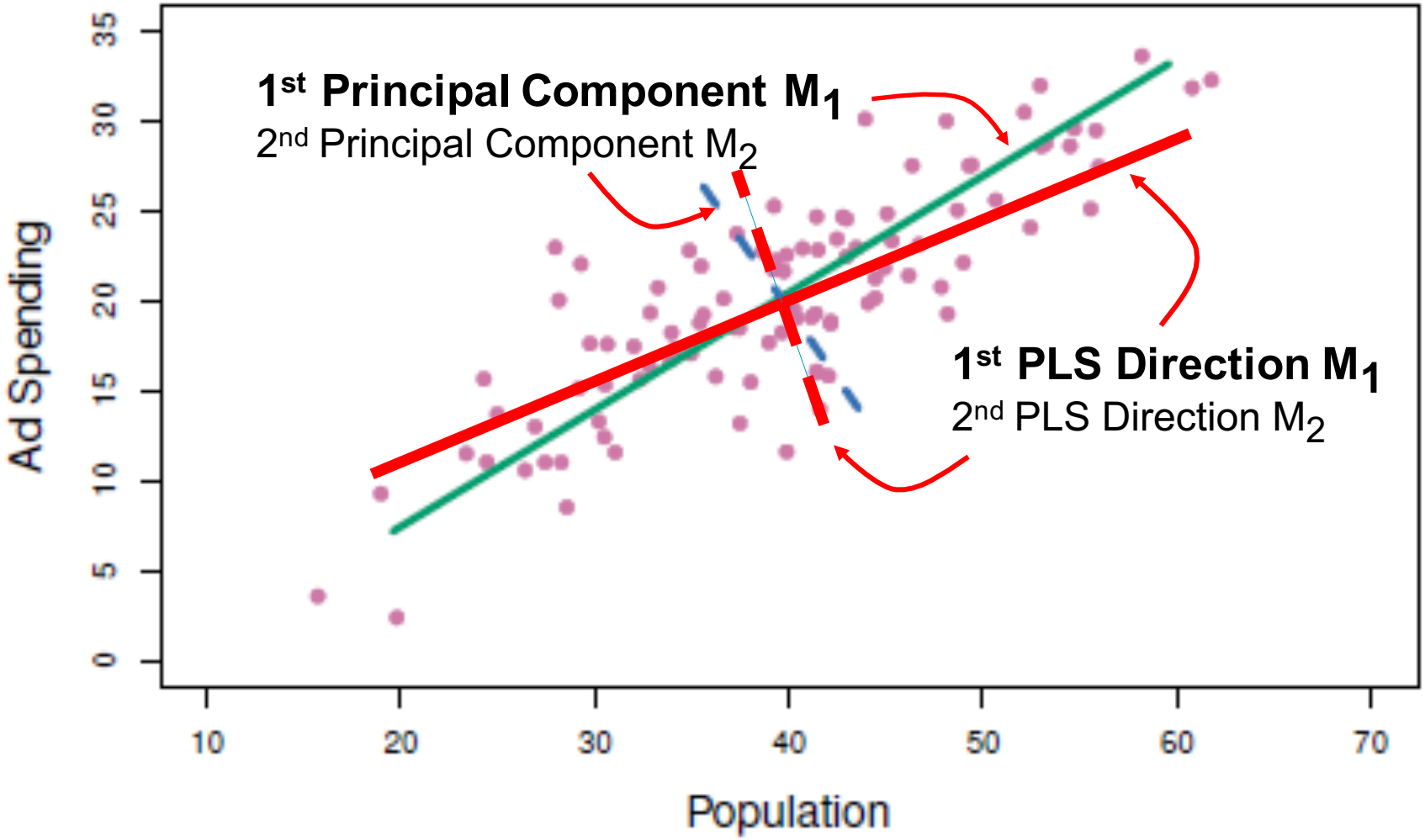
Partial Least Squares (PLS) Regression



Partial Least Squares (PLS): Intuition

- With **PCR**, the $X_1, X_2, \ldots\ldots X_P$ variables are transformed into $M_1, M_2, \ldots\ldots M_P$ components in an **"unsupervised"** way
- That is, the independent variable dimensions are **rotated** to find directions in which the data exhibits highest variance.
- This is an **"unsupervised"** method because the **outcome** variable Y is not taken into account when doing PCA
- While PCR does well in general, there is **no guarantee** the first few M **components** will be the **best directions** to predict **Y**
- In contrast, PLS is a "supervised" method
- Like PCR, **PLS** is a **dimension reduction** method, but unlike PCR, PLS does **further rotation** of the dimensions to maximize the **correlation** with **Y**
- In sum, PLS attempts to find directions that **not only explain** the **predictors**, but also the **outcome** variable
- It does this by placing stronger weight on variables that are more strongly correlated with Y

PLS: Illustration



Components **M_1** and **M_2** are **not**

PLS Regression

- In **PLS** the components are **no longer PC's** because they deviate from the directions of higher variance, so they are simply called first, second, etc. **"directions"**.
- Like with PCR, PLS starts with **OLS** (X's are **standardized**):

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

- And finds **M "directions"** that are linear combinations of the X'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)
- But the identification of the first direction $\mathbf{M_1}$ starts by placing more **weight** on the variable that has the strongest **correlation** with **Y**
- Like with PCR, **M** is a **"tuning"** paramerer.
- As with OLS and PCR, in **PLS** as **M** ↑ → **Bias** ↓ and **Variance** ↑
- Because **PLS** is a **supervised** method, the coefficients are **less biased** but **more variance** than **PCR**
- Again, **cross-validation** is the best selector to find the optimal **M** and to compare with other modeling approaches

Tips

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

The `pslr()` function syntax is identical to the `pcr()` function syntax

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

KOGOD SCHOOL
of
BUSINESS