Multivariate Analysis
Canonical Correlation Analysis

Shaobo Jin

Department of Mathematics

Intended Learning Outcome

Through this chapter, you should be able to

1. perform canonical correlation analysis,
2. perform partial least squares regression.

Motivation

- PCA performs dimension reduction for one set of p variables, while extracting as much information as possible.
- Now suppose that we have a $p \times 1$ vector of variables $\boldsymbol{X}^{(1)}$, and a $q \times 1$ vector of variables $\boldsymbol{X}^{(2)}$, measured on the same individual. We want to summarize the relationships between two vectors. '
  - e.g., you have p variables ($\boldsymbol{X}^{(1)}$) to measure an individual's motivations for watching online videos, and q variables ($\boldsymbol{X}^{(2)}$) to measure how the individual access the online videos.
  - pq pairwise scatter plots.
  - Canonical correlation analysis (CCA) focuses on linear combinations of variables such that much fewer plots are needed.

Task

Let $\boldsymbol{X}^{(1)}$ be a $p \times 1$ random vector, and $\boldsymbol{X}^{(2)}$ be a $q \times 1$ random vector. We assume $p \leq q$ with loss of generality. Let

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix},$$

where

$$\mathbb{E}\left(\boldsymbol{X}\right) = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}, \qquad \operatorname{cov}\left(\boldsymbol{X}\right) = \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

The main task of CCA is to find linear combinations

$$U = \boldsymbol{a}^T \boldsymbol{X}^{(1)}, \qquad V = \boldsymbol{b}^T \boldsymbol{X}^{(2)},$$

such that $\operatorname{corr}\left(U, V\right)$ is maximized.

Correlation Coefficient

By Result 4.2,

$$\begin{aligned}
\operatorname{cov}\left(\begin{bmatrix} U \\ V \end{bmatrix}\right) &= \operatorname{cov}\left(\begin{bmatrix} \boldsymbol{a}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{b}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \end{bmatrix}\right) \\
&= \begin{bmatrix} \boldsymbol{a}^T & \mathbf{0} \\ \mathbf{0} & \boldsymbol{b}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{a} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{b} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{a}^T \boldsymbol{\Sigma}_{11} \boldsymbol{a} & \boldsymbol{a}^T \boldsymbol{\Sigma}_{12} \boldsymbol{b} \\ \boldsymbol{b}^T \boldsymbol{\Sigma}_{12}^T \boldsymbol{a} & \boldsymbol{b}^T \boldsymbol{\Sigma}_{22} \boldsymbol{b} \end{bmatrix}.
\end{aligned}$$

We shall seek coefficient vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ such that

$$\operatorname{corr}(U, V) = \frac{\boldsymbol{a}^T \boldsymbol{\Sigma}_{12} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^T \boldsymbol{\Sigma}_{11} \boldsymbol{a}} \sqrt{\boldsymbol{b}^T \boldsymbol{\Sigma}_{22} \boldsymbol{b}}}$$

is as large as possible.

Restriction

Note that

$$\begin{aligned} \operatorname{corr}(U, V) &= \frac{\boldsymbol{a}^T \boldsymbol{\Sigma}_{12} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^T \boldsymbol{\Sigma}_{11} \boldsymbol{a}} \sqrt{\boldsymbol{b}^T \boldsymbol{\Sigma}_{22} \boldsymbol{b}}} \\ s &= \frac{(c\boldsymbol{a})^T \boldsymbol{\Sigma}_{12} (c\boldsymbol{b})}{\sqrt{(c\boldsymbol{a})^T \boldsymbol{\Sigma}_{11} (c\boldsymbol{a})} \sqrt{(c\boldsymbol{b})^T \boldsymbol{\Sigma}_{22} (c\boldsymbol{b})}}. \end{aligned}$$

Hence, we need to set the scales of $\boldsymbol{a}$ and $\boldsymbol{b}$. One option is

$$\begin{aligned} \boldsymbol{a}^T \boldsymbol{\Sigma}_{11} \boldsymbol{a} &= 1, \\ \boldsymbol{b}^T \boldsymbol{\Sigma}_{22} \boldsymbol{b} &= 1. \end{aligned}$$

Canonical Variates

1. The first canonical variate pair is the pair of linear combinations U_1 and V_1 having unit variances which maximizes the correlation $\text{corr}\left(U_1, V_1\right)$ (first canonical correlation).
2. The second canonical variate pair is the pair of linear combinations U_2 and V_2 having unit variances which maximizes the correlation $\text{corr}\left(U_2, V_2\right)$ (second canonical correlation) among all choices that are uncorrelated with the first pair of canonical variables.
3. The kth canonical variate pair is the pair of linear combinations U_k and V_k having unit variances which maximizes the correlation $\text{corr}\left(U_k, V_k\right)$ (kth canonical correlation) among all choices that are uncorrelated with the previous $k-1$ pairs of canonical variables.

A Useful Lemma

Lemma

If λ is an eigenvalue of $\boldsymbol{AB}$ with corresponding eigenvector $\boldsymbol{x}$, i.e.,

$$\boldsymbol{ABx} = \lambda \boldsymbol{x}.$$

Then, λ is an eigenvalue of $\boldsymbol{BA}$ with corresponding eigenvector $\boldsymbol{Bx}$, since

$$\boldsymbol{BABx} = \lambda \boldsymbol{Bx}.$$

Find Canonical Variates

Result 10.1

Suppose $p \leq q$ and let the random vectors $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ have a full rank covariance matrix $\boldsymbol{\Sigma}$. Then the kth canonical correlation is ρ_k^*, and the kth canonical variate pair is attained by

$$\begin{aligned} U_k &= \left(\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_k\right)^T \boldsymbol{X}^{(1)}, \\ V_k &= \left(\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{f}_k\right)^T \boldsymbol{X}^{(2)}. \end{aligned}$$

Here $\rho_1^{*2} \geq \rho_2^{*2} \geq \cdots \geq \rho_p^{*2}$ are the eigenvalues of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1/2}$, and $\boldsymbol{e}_k$ are the associated eigenvectors. Each $\boldsymbol{f}_k$ is proportional to $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_k$.

Unit Variance and Zero Correlation

The canonical variates have the properties

$$\begin{aligned} \operatorname{var}(U_k) &= \operatorname{var}(V_k) = 1, \\ \operatorname{cov}(U_k, U_\ell) &= 0, \quad k \neq \ell, \\ \operatorname{cov}(V_k, V_\ell) &= 0, \quad k \neq \ell, \\ \operatorname{cov}(U_k, V_\ell) &= 0, \quad k \neq \ell. \end{aligned}$$

for all k, $\ell = 1, 2, ..., p$.
If we apply PCA to $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ separately, then we only have

$$\begin{aligned} \operatorname{cov}(U_k, U_\ell) &= 0, \quad k \neq \ell, \\ \operatorname{cov}(V_k, V_\ell) &= 0, \quad k \neq \ell. \end{aligned}$$

Scale Invariant: Canonical Correlation

Suppose that we change the scale as $\boldsymbol{Z}^{(1)} = \boldsymbol{C}_1 \boldsymbol{X}^{(1)} + \boldsymbol{d}_1$ and $\boldsymbol{Z}^{(2)} = \boldsymbol{C}_2 \boldsymbol{X}^{(2)} + \boldsymbol{d}_2$, where $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ are invertible. Then,

$$\operatorname{cov}\left(\begin{bmatrix}\boldsymbol{Z}^{(1)} \\ \boldsymbol{Z}^{(2)}\end{bmatrix}\right) = \begin{bmatrix}\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T & \boldsymbol{C}_1\boldsymbol{\Sigma}_{12}\boldsymbol{C}_2^T \\ \boldsymbol{C}_2\boldsymbol{\Sigma}_{12}^T\boldsymbol{C}_1^T & \boldsymbol{C}_2\boldsymbol{\Sigma}_{22}\boldsymbol{C}_2^T\end{bmatrix}.$$

The kth canonical correlation is r_k^*, the eigenvalues of

$$\begin{aligned} & \left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2} \boldsymbol{C}_1\boldsymbol{\Sigma}_{12}\boldsymbol{C}_2^T \left(\boldsymbol{C}_2\boldsymbol{\Sigma}_{22}\boldsymbol{C}_2^T\right)^{-1} \left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{12}\boldsymbol{C}_2^T\right)^T \left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2} \\ = & \left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2} \boldsymbol{C}_1\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{C}_1^T \left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}. \end{aligned}$$

They are the same as the nonzero eigenvalues of

$$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1/2}.$$

Scale Invariant: Coefficient Vector

Let $\boldsymbol{e}_k^{(Z)}$ satisfies

$$\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{C}_1\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{C}_1^T\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2} = \rho_k^*\boldsymbol{e}_k^{(Z)}.$$

Then,

$$\begin{aligned}
\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1}\boldsymbol{C}_1\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{C}_1^T\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{e}_k^{(Z)} &= \rho_k^*\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{e}_k^{(Z)}\\
\Rightarrow \quad \left(\boldsymbol{C}_1^T\right)^{-1}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{C}_1^T\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{e}_k^{(Z)} &= \rho_k^*\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{e}_k^{(Z)}\\
\Rightarrow \quad \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}\times\boldsymbol{C}_1^T\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{e}_k^{(Z)} &= \rho_k^*\boldsymbol{C}_1^T\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{e}_k^{(Z)}.
\end{aligned}$$

This means that

$$\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{e}_k = \boldsymbol{C}_1^T\left(\boldsymbol{C}_1\boldsymbol{\Sigma}_{11}\boldsymbol{C}_1^T\right)^{-1/2}\boldsymbol{e}_k^{(Z)}.$$

Canonical Correlation As Correlation Bound

Note that

$$\left|\operatorname{cor}\left(X_i^{(1)}, X_j^{(j)}\right)\right| \quad \leq \quad \max_{a,b}\left|\operatorname{corr}\left(\boldsymbol{a}^T\boldsymbol{X}^{(1)}, \boldsymbol{b}^T\boldsymbol{X}^{(2)}\right)\right| = \rho_1^*.$$

Hence, the first canonical correlation is no lower than the absolute value of the correlation between any $X_i^{(1)}$ and $X_j^{(j)}$.

Proportion of Explained Variance

In fact,

$$\begin{aligned} \max_{b} \operatorname{corr}\left(U_k, \boldsymbol{b}^T \boldsymbol{X}^{(2)}\right) &= \rho_k^*, \\ \max_{a} \operatorname{corr}\left(\boldsymbol{a}^T \boldsymbol{X}^{(1)}, V_k\right) &= \rho_k^*, \end{aligned}$$

for any k.
The kth squared canonical correlation ρ_k^{*2} is

- the proportion of the variance of canonical variate U_k explained by the set $\boldsymbol{X}^{(2)}$,
- the proportion of the variance of canonical variate V_k explained by the set $\boldsymbol{X}^{(1)}$.

Hence, ρ_k^{*2} is the shared variance between $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$.

Sample Canonical Variate Pair

A random sample of n observations is assembled into the $n \times (p+q)$ data matrix $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}^{(1)} & \boldsymbol{X}^{(2)} \end{bmatrix}$.

1. The first sample canonical variate pair is the pair of linear combinations $\hat{U}_1$ and $\hat{V}_1$ having unit variances which maximizes the sample correlation $r_{\hat{U}_1,\hat{V}_1}$.
2. The second sample canonical variate pair is the pair of linear combinations $\hat{U}_2$ and $\hat{V}_2$ having unit variances which maximizes the sample correlation $r_{\hat{U}_2,\hat{V}_2}$ among all choices that are uncorrelated with the first sample canonical variate pair.
3. The kth sample canonical variate pair is the pair of linear combinations $\hat{U}_k$ and $\hat{V}_k$ having unit variances which maximizes the sample correlation $r_{\hat{U}_k,\hat{V}_k}$ among all choices that are uncorrelated with the previous $k-1$ sample canonical variate pairs.

Sample Correlation

Let

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{12}^T & \boldsymbol{S}_{22} \end{bmatrix}$$

be the sample covariance matrix of $\boldsymbol{X}$. The sample correlation between $\boldsymbol{a}^T\boldsymbol{X}^{(1)}$ and $\boldsymbol{b}^T\boldsymbol{X}^{(2)}$ is

$$\frac{\boldsymbol{a}^T\boldsymbol{S}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T\boldsymbol{S}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}^T\boldsymbol{S}_{22}\boldsymbol{b}}}.$$

The population correlation is

$$\frac{\boldsymbol{a}^T\boldsymbol{\Sigma}_{12}\boldsymbol{b}}{\sqrt{\boldsymbol{a}^T\boldsymbol{\Sigma}_{11}\boldsymbol{a}}\sqrt{\boldsymbol{b}^T\boldsymbol{\Sigma}_{22}\boldsymbol{b}}}.$$

Find Sample Canonical Variate Pair

Result 10.2

Let $\hat{\rho}_1^{*2} \geq \hat{\rho}_2^{*2} \geq \cdots \geq \hat{\rho}_p^{*2}$ be the eigenvalues of $\boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2}$ with corresponding eiegnvectors $\hat{\boldsymbol{e}}_1$, ..., $\hat{\boldsymbol{e}}_p$, where $p \leq q$. Let $\hat{\boldsymbol{f}}_1$, ..., $\hat{\boldsymbol{f}}_p$ be the eigenvectors of $\boldsymbol{S}_{22}^{-1/2}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1/2}$ where the first p eigenvectors may be obtained from $\hat{\boldsymbol{f}}_k = \hat{\rho}_k^{*\,-1}\boldsymbol{S}_{22}^{-1/2}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1}\hat{\boldsymbol{e}}_k$, $k = 1, 2, ..., p$. Then the kth sample canonical variate pair is

$$\hat{U}_k = \left(\boldsymbol{S}_{11}^{-1/2}\hat{\boldsymbol{e}}_k\right)^T \boldsymbol{x}^{(1)}, \qquad \hat{V}_k = \left(\boldsymbol{S}_{22}^{-1/2}\hat{\boldsymbol{f}}_k\right)^T \boldsymbol{x}^{(2)}.$$

The kth sample canonical correlation is $\hat{\rho}_k^*$.

Unit Variance and Zero Correlation

The sample canonical variates have the properties

$$
\begin{aligned}
S_{\hat{U}_k} &= S_{\hat{V}_k} = 1, \\
r_{\hat{U}_k,\hat{U}_\ell} &= 0, \quad k \neq \ell, \\
r_{\hat{V}_k,\hat{V}_\ell} &= 0, \quad k \neq \ell, \\
r_{\hat{U}_k,\hat{V}_\ell} &= 0, \quad k \neq \ell,
\end{aligned}
$$

for all k, $\ell = 1, 2, ..., p$.
If we apply PCA to $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ separately, then we only have

$$
\begin{aligned}
r_{\hat{U}_k,\hat{U}_\ell} &= 0, \quad k \neq \ell, \\
r_{\hat{V}_k,\hat{V}_\ell} &= 0, \quad k \neq \ell.
\end{aligned}
$$

Special Case: $p = 1$

Consider a special case where $p = 1$. We denote $Y = X^{(1)}$ and $\boldsymbol{Z} = \boldsymbol{X}^{(2)}$. The matrix $\boldsymbol{S}_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{12}^T\boldsymbol{S}_{11}^{-1/2}$ reduces to a scalar with eigenvalue itself and eigenvector 1. Then, the canonical variate pair is attained by

$$\begin{aligned} V &= \left(\boldsymbol{S}_{22}^{-1/2}\boldsymbol{f}\right)^T\boldsymbol{Z} = \left(\boldsymbol{S}_{22}^{-1/2} \times |\rho^*|\, \boldsymbol{S}_{22}^{-1/2}\boldsymbol{S}_{12}^T S_{11}^{-1/2}\hat{e}\right)^T\boldsymbol{Z} \\ &= S_{11}^{-1}\sqrt{\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{12}^T}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{Z}, \end{aligned}$$

where $\hat{\rho}^{*2} = S_{11}^{-1/2}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{12}^T S_{11}^{-1/2} = S_{11}^{-1}\boldsymbol{S}_{12}\boldsymbol{S}_{22}^{-1}\boldsymbol{S}_{12}^T$ and $\hat{e} = 1$.

If we apply classic linear regression (regress Y on $\boldsymbol{Z}$), the OLS fitted model is

$$\hat{Y} = \underbrace{\boldsymbol{y}^T\boldsymbol{Z}_D}_{\propto \boldsymbol{S}_{12}}\underbrace{\left(\boldsymbol{Z}_D^T\boldsymbol{Z}_D\right)^{-1}}_{\propto \boldsymbol{S}_{22}^{-1}}\boldsymbol{Z},$$

where $\boldsymbol{Z}_D$ is the demeaned data matrix of $\boldsymbol{Z}$ and $\boldsymbol{y}$ is the demeaned response vector of Y.

CCA Versus Regression: Seemingly Different

```r
## Canonical Correlation Analysis
CC <- cc(X = as.matrix(Data[, "Y"]),
         Y = as.matrix(Data[, c("Z1", "Z2", "Z3")]))
CC$ycoef

##         [,1]
## Z1 0.6444941
## Z2 0.3009713
## Z3 0.3615434

## Classic Linear Regression
LM <- lm(Y ~ Z1 + Z2 + Z3, data = Data)
coef(LM)

## (Intercept)          Z1          Z2          Z3
##   0.1486136   0.2558822   0.1194941   0.1435429
```

More Outputs of Linear Regression

```
summary(LM)

##
## Call:
## lm(formula = Y ~ Z1 + Z2 + Z3, data = Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.30639 -0.61481  0.05427  0.56709  1.90861
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.14861    0.09000   1.651   0.1019
## Z1           0.25588    0.09957   2.570   0.0117 *
## Z2           0.11949    0.09509   1.257   0.2119
## Z3           0.14354    0.09993   1.436   0.1541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8756 on 96 degrees of freedom
## Multiple R-squared:  0.1749,Adjusted R-squared:  0.1492
## F-statistic: 6.785 on 3 and 96 DF,  p-value: 0.0003378
```

Almost Same Thing

```r
## Just Scaling
c(CC$ycoef / coef(LM)[-1])

## [1] 2.518714 2.518714 2.518714

## Canonical Correlation is related
c(CC$cor ^ 2, summary(LM)$r.squared)

## [1] 0.1749436 0.1749436
```

Alternative to Multivariate Multiple Regression

Suppose that we have p response variables and q covariates that we can use to model the response variable.

1. One approach is to use multivariate multiple regression (Chapter 7).
2. An alternative approach is partial least squares regression (PLS regression).

A little history:

- Karl Gustav Jöreskog popularized factor analysis. He had professorship at Uppsala 1971–2000.
- Herman Wold developed PLS regression. He had professorship at Uppsala 1942–1970.
- Jöreskog was a student of Wold at Uppsala.

Maximize Covariance

PLS regression is an alternative to multivariate multiple regression if

- n is relatively small comparing to p or q,
- correlation between covariates are high (multicollinearity).

Suppose that we have demeaned the p response variables and q covariates, such that the sample mean of each variable is 0. We want to find linear combinations $T = \boldsymbol{a}^T \boldsymbol{X}$ and $U = \boldsymbol{b}^T \boldsymbol{Y}$ such that the covariance between T_k and U is maximized. That is,

$$\max_{a,b} \operatorname{cov}\left(\boldsymbol{a}^T \boldsymbol{X}, \boldsymbol{b}^T \boldsymbol{Y}\right)$$
$$\text{s.t. } \boldsymbol{a}^T \boldsymbol{a} = \boldsymbol{b}^T \boldsymbol{b} = 1.$$

This is similar to the objective of CCA.

Eigenvalue and Eigenvector

Since the variables are demeaned, the sample covariance matrix between $\boldsymbol{X}_{n\times q}$ and $\boldsymbol{Y}_{n\times p}$ can be computed by $n^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ or $(n-1)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Then,

$$\max_{a,b} \boldsymbol{a}^T\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{b} \qquad \text{s.t. } \boldsymbol{a}^T\boldsymbol{a} = \boldsymbol{b}^T\boldsymbol{b} = 1.$$

Hence, $\boldsymbol{a}$ is the eigenvector of $\boldsymbol{X}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{X}$ corresponding to the largest eigenvalue, and $\boldsymbol{b} \propto \boldsymbol{Y}^T\boldsymbol{X}\boldsymbol{a}$.

1. We regress $\boldsymbol{X}$ on $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{a}$, a $n \times 1$ vector. The OLS estimator of regression coefficients is $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{t}^T\boldsymbol{t}\right)^{-1}\boldsymbol{t}^T\boldsymbol{X}$.
2. We regress $\boldsymbol{Y}$ also on $\boldsymbol{t}$. The OLS estimator of regression coefficients is $\hat{\boldsymbol{\gamma}} = \left(\boldsymbol{t}^T\boldsymbol{t}\right)^{-1}\boldsymbol{t}^T\boldsymbol{Y}$.

PLS Regression

From the regression model of $\boldsymbol{Y}$ on $\boldsymbol{t}$, our fitted model for a new value t_0 satisfies

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\gamma}} t_0.$$

Since $t_0 = \boldsymbol{a}^T \boldsymbol{x}_0$, then

$$\hat{\boldsymbol{y}} = \hat{\boldsymbol{\gamma}} \boldsymbol{a}^T \boldsymbol{x}_0,$$

which is the PLS regression of $\boldsymbol{y}$ on $\boldsymbol{x}$.

More Components

So far we have only used one pair of linear combinations. Similar to CCA, we can extract more components.

- When we regress $\boldsymbol{X}$ on $\boldsymbol{t} = \boldsymbol{X}\boldsymbol{a}$, the residual matrix is

$$\boldsymbol{E}_X = \boldsymbol{X} - \boldsymbol{t}\left(\boldsymbol{t}^T\boldsymbol{t}\right)^{-1}\boldsymbol{t}^T\boldsymbol{X}.$$

- We regress $\boldsymbol{Y}$ also on $\boldsymbol{t}$, the residual matrix is

$$\boldsymbol{E}_Y = \boldsymbol{Y} - \boldsymbol{t}\left(\boldsymbol{t}^T\boldsymbol{t}\right)^{-1}\boldsymbol{t}^T\boldsymbol{Y}.$$

- We treat $\boldsymbol{E}_X$ and $\boldsymbol{E}_Y$ as $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively in a new iteration of PLS regression.