Multiple Linear Regression

Examples

- 晶圓(Wafer)良率或不良率～多個機台+機台參數
- 螢幕Mura～機台+材料面板瑕疵
- 汽車保養頻率～引擎+組裝+動力系統
- 眼科疾病～年紀+糖尿病+血壓+家族病史
- GDP～人口數+出口+進口+…

Multiple linear regression model

$$Y \sim \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \cdots$$

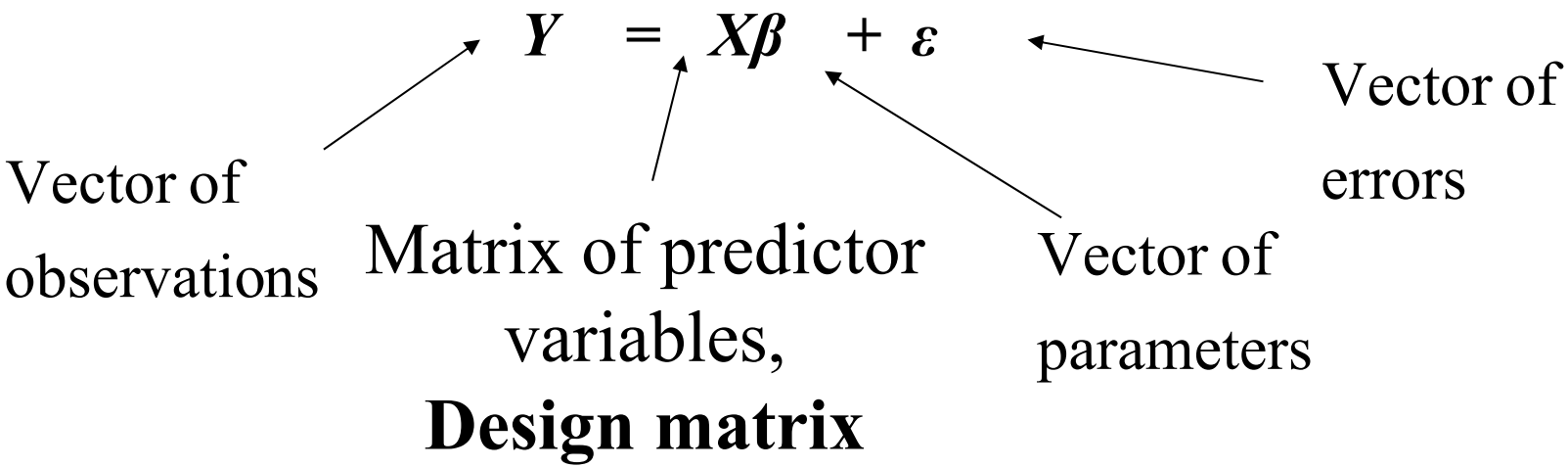
$$Y \sim \beta_0 + \beta_1 X_1 + \beta_2 X_2 = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

$$\left(\text{e.g. } X_2 = X_1^2\right)$$

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, \cdots, n$$

- Condition(A1). $E(\varepsilon_i) = 0, \ \ i = 1, \cdots, n.$
- Condition(A2). $Var(\varepsilon_i) = \sigma^2, \ \ i = 1, \cdots, n.$
- Condition(A3). $Cov(\varepsilon_i, \varepsilon_j) = 0, \ \text{for } i \neq j.$

Regression in Matrix Terms



$$\mathbf{Y} = \underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

RSS (SSE) and LSE

$$RSS(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \left(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \right) \right)^2$$

$$= \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = \left(\boldsymbol{y} - \boldsymbol{X\beta} \right)^T \left(\boldsymbol{y} - \boldsymbol{X\beta} \right) = \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \left(\boldsymbol{X}^T \boldsymbol{y} \right) + \boldsymbol{\beta}^T \left(\boldsymbol{X}^T \boldsymbol{X} \right) \boldsymbol{\beta}$$

Derivatives the RSS and let it be zero, we have

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

- $E\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \boldsymbol{\beta}$

- $Cov\left(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}\right) = \left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\sigma^2$

- $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$

$\boldsymbol{H}$: projection matrix or hat matrix

$$h_{ii} = \boldsymbol{x}_i^T \left(\boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{x}_i^T$$

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \left(\boldsymbol{I} - \boldsymbol{H} \right) \boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}$$

Notes: $\boldsymbol{H}$ and $\boldsymbol{M}$ are symmetric and idempotent.

$$\hat{\boldsymbol{\varepsilon}}^{T}\,\boldsymbol{X} \;=\; \boldsymbol{0}$$

$$\hat{\boldsymbol{\varepsilon}}^{T}\,\boldsymbol{1} \;=\; \boldsymbol{0}$$

$$\hat{\boldsymbol{\varepsilon}}^{T}\,\hat{\boldsymbol{y}} \;=\; \boldsymbol{0}$$

$$\hat{\sigma}^{2} \;=\; s^{2} \;=\; \frac{\hat{\boldsymbol{\varepsilon}}^{T}\,\hat{\boldsymbol{\varepsilon}}}{n - (p + 1)}$$

$$SST(n-1) = SSR(p) + SSE(n-(p+1))$$

$$\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$$

$$\Leftrightarrow \left(\boldsymbol{y} - \boldsymbol{I}\bar{\boldsymbol{y}}\right)^{\boldsymbol{T}}\left(\boldsymbol{y} - \boldsymbol{I}\bar{\boldsymbol{y}}\right) = \left(\hat{\boldsymbol{y}} - \boldsymbol{I}\bar{\boldsymbol{y}}\right)^{\boldsymbol{T}}\left(\hat{\boldsymbol{y}} - \boldsymbol{I}\bar{\boldsymbol{y}}\right) + \hat{\boldsymbol{\varepsilon}}^{\boldsymbol{T}}\hat{\boldsymbol{\varepsilon}}$$

$$\Leftrightarrow \boldsymbol{y}^{\boldsymbol{T}}\left(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^{\boldsymbol{T}}}{\boldsymbol{n}}\right)\boldsymbol{y} = \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{H}^{\boldsymbol{T}}\left(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^{\boldsymbol{T}}}{\boldsymbol{n}}\right)\boldsymbol{H}\boldsymbol{y} + \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{M}\boldsymbol{y}$$

$$\Leftrightarrow \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{L}\boldsymbol{y} = \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{L}\boldsymbol{H}\boldsymbol{y} + \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{M}\boldsymbol{y}$$

$$R^2 = \frac{SSR}{SST} = \frac{\text{Variability explained by the model}}{\text{Total sample variability}} = \frac{\boldsymbol{y}^T \boldsymbol{H}^T \boldsymbol{L} \boldsymbol{H} \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{L} \boldsymbol{y}}$$

$$= 1 - \frac{\text{Unexplained(error)}}{\text{Total sample variability}} = 1 - \frac{\boldsymbol{y}^T \boldsymbol{M} \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{L} \boldsymbol{y}}$$

$$= 1 - \frac{SSE}{SST}$$

$$0 \le R^2 \le 1$$

ANOVA table

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \text{ versus } H_1 : \beta_j \neq 0 \text{ for some } j$$

Source of variation (Source)	Sum of squares (SS)	Degrees of freedom (df)	Mean square (MS)	F value (F)	P-value
Regression	SSR	p	$MSR = SSR / p$	$F = MSR / MSE$	$P\left(F_{p,n-(p+1)} > F\right)$
Error	SSE	n-(p+1)	$MSE = SSE / \left(n-(p+1)\right)$		
Total	SST	n-1			

Testing whether a specified subset of the predictors have coefficients equal to 0

Recall:

$H_0 : \beta_1 = \cdots = \beta_p = 0$ versus $H_1 : \beta_j \neq 0,$ for some $j = 1, \ldots, p$

i.e. $Y_i = \beta_0 + \varepsilon_i$, for $i = 1, \ldots, n$ and

$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$ for $i = 1, \ldots, n$

and for at least one of $\beta_j \neq 0$, $j = 1, \ldots, p$

$$F = \frac{SSR / p}{SSE / \left(n - \left(p + 1\right)\right)} \sim F\left(p, n - \left(p + 1\right)\right)$$

Reject region $C = \left\{F > F_\alpha\left(p, n - \left(p + 1\right)\right)\right\}$

$\mathrm{H}_0 : \beta_{m+1} = \cdots = \beta_p = 0$ versus $\mathrm{H}_1 : \beta_j \neq 0,$ for some $j = m+1, \ldots, p$

two nested models, H_0 : reduced model versus H_1 : full model

$$\mathrm{H}_0 : \; Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \varepsilon_i, \text{ for } i = 1, \ldots, n \text{ and}$$

$$\mathrm{H}_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_m x_{im} + \beta_{m+1} x_{im+1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

for $i = 1, \ldots, n$ and for at least one of $\beta_j \neq 0,\ j = m+1, \ldots, p$

$$\text{(partial) } \mathrm{F} = \frac{\left(SSR_{n-(m+1)} - SSR_{n-(p+1)} \right) / \left(p - m \right)}{SSE \,/ \left(n - \left(p + 1 \right) \right)} \sim F\left(p - m, n - \left(p + 1 \right) \right)$$

$$\text{Reject region } C = \left\{ \mathrm{F} > F_\alpha \left(p - m, n - \left(p + 1 \right) \right) \right\}$$

$H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0,$ for $j \in \{1, \ldots, p\}$

two nested models, H_0 : reduced (restricted model) model versus

H_1 : full (unrestricted model) model

$$H_0 : Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{i\,j-1} + \beta_{j+1} x_{i\,j+1} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

for $i = 1, \ldots, n,\ j \in \{1, \ldots, p\}$ and

$$H_1 : Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \text{ for } i = 1, \ldots, n$$

$$\text{(partial) } F = \frac{\left(SSR_{n-p} - SSR_{n-(p+1)}\right)}{SSE / \left(n - (p+1)\right)} \sim F\left(1, n - (p+1)\right)$$

$$\text{Reject region } C = \left\{F > F_\alpha\left(1, n - (p+1)\right)\right\}$$

Sequential F test statistics

For $Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i,\ i = 1, \ldots, n,$ consider

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0. \Rightarrow F_{1,n-2}$

Next, for $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i,\ i = 1, \ldots, n,$ consider

$H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0. \Rightarrow F_{1,n-3}$

$\vdots$

For $Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \varepsilon_i,\ i = 1, \ldots, n,$ consider

$H_0 : \beta_p = 0$ versus $H_1 : \beta_p \neq 0. \Rightarrow F_{1,n-p-1}$

Then, the nested F test statistic, $F_{1,n-2}, F_{1,n-3}, \ldots, F_{1,n-p-1},$ are called seqential F test statistic.

Dummy variables

Binary variable

- sex (female or male)
- disease (yes or no)
- master degree (yes or no)
- …

Parallel regression lines

$$Y_i = \beta_0 + \beta_1 x_i + \delta d_i + \varepsilon_i, \quad i = 1, \cdots, n$$

x_i : continuous covariate

d_i : binary covariate

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i & \text{if } d_i = 0 \\ (\beta_0 + \delta) + \beta_1 x_i + \varepsilon_i & \text{if } d_i = 1 \end{cases},$$

$d_i = 0$: base group

Regression lines with equal intercepts but different slopes

$$Y_i = \beta_0 + \beta_1 x_i + \delta' d_i x_i + \varepsilon_i, \quad i = 1, \cdots, n$$

x_i : continuous covariate

d_i : binary covariate

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i & \text{if } d_i = 0 \\ \beta_0 + (\beta_1 + \delta') x_i + \varepsilon_i & \text{if } d_i = 1 \end{cases},$$

$d_i = 0$: base group

Unrelated regression lines

$$Y_i = \beta_0 + \beta_1 x_i + \delta d_i + \delta' d_i x_i + \varepsilon_i, \quad i = 1, \cdots, n$$

x_i : continuous covariate

d_i : binary covariate

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i & \text{if } d_i = 0 \\ (\beta_0 + \delta) + (\beta_1 + \delta') x_i + \varepsilon_i & \text{if } d_i = 1 \end{cases},$$

$d_i = 0$: base group

Dummy variables (cont’)

Categorical variable

- salary level (150+, 100~150, 50~100,< 50)
- education level(phd, master, bachelor, other)
- ethnic ("Aethiopian or Black", "Caucasian or White", "Mongolian or Yellow", "American or Red" and "Malayan or Brown")
- …

Categorical variable: four levels

	d1	d2	d3
C1	0	0	0
C2	1	0	0
C3	0	1	0
C4	0	0	1

Parallel regression lines

$$Y_i = \beta_0 + \beta_1 x_i + \delta_1 d_{i1} + \delta_2 d_{i2} + \delta_3 d_{i3} + \varepsilon_i, \quad i = 1, \cdots, n$$

x_i : continuous covariate

d_i : dummy variables for the categorical covariate

$$Y_i = \begin{cases} \beta_0 + \beta_1 x_i + \varepsilon_i & \text{if C1} \\ (\beta_0 + \delta_1) + \beta_1 x_i + \varepsilon_i & \text{if C2} \\ (\beta_0 + \delta_2) + \beta_1 x_i + \varepsilon_i & \text{if C3} \\ (\beta_0 + \delta_3) + \beta_1 x_i + \varepsilon_i & \text{if C4} \end{cases}$$

$(d_1, d_2, d_3) = (0, 0, 0)$: base group

Categorical variable: four levels

	d1	d2	d3
C1	0	0	0
C2	1	0	0
C3	1	1	0
C4	1	1	1

Collinearity in linear model with the intercept term

	d1	d2	d3	D4
C1	1	0	0	0
C2	0	1	0	0
C3	0	0	1	0
C4	0	0	0	1

Multi-collinearity

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x}_1 + \cdots + \beta_p \boldsymbol{x}_p + \boldsymbol{\varepsilon}.$$

What is multicollinearity?

If there is a set of $c_1, \ldots, c_p$ (not all zero) for which

$$\sum_{j=1}^{p} c_j \boldsymbol{x}_j \cong \boldsymbol{0}.$$

What is the impact of multicollinearity?

The LSE, $\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}$, will not be stable.

Consider the following special case,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \quad i = 1, \cdots, n,$$

where $Var\left(\varepsilon_i\right) = \sigma^2$ and all the ε_i are uncorrelated.

Assume that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ have been scaled so that

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2 \cdot \sum_{i=1}^{n}\left(x_{i1} - \bar{x}_1\right)^2}{1 - r_1^2} = \frac{\sigma^2}{1 - r_1^2}.$$

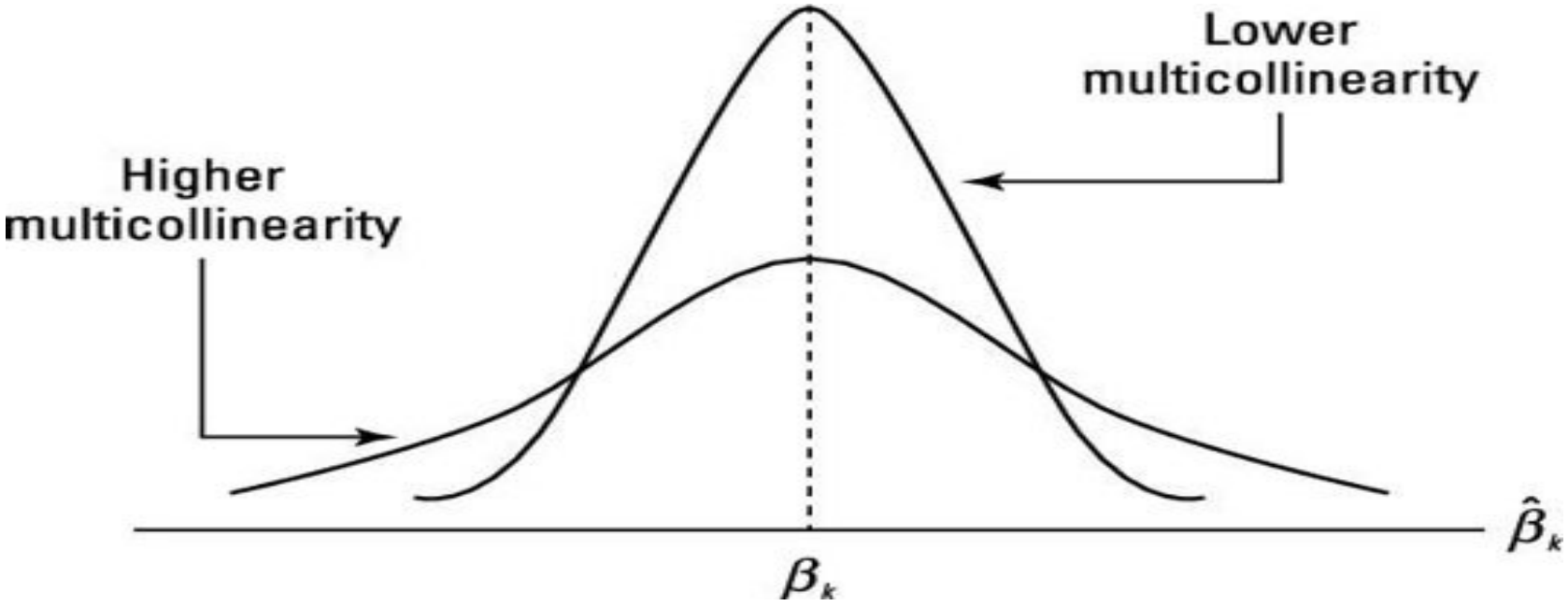
$1-r_1^2$ is the residual mean square error (RMSE) obtained from regressing $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. In particular,

$$r_1^2 = \frac{\left(\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_1\right)\left(x_{i2}-\bar{x}_2\right)\right)^2}{\sum_{i=1}^{n}\left(x_{i1}-\bar{x}_2\right)^2}.$$

- When $r_1^2 = 0, Var\left(\hat{\beta}_1\right)$ has its minimum value.
- As $r_1^2 \to 1$, $Var\left(\hat{\beta}_1\right)$ will get to be arbitrarily large.

Variance inflation factors (VIFs)

$\text{VIFs} = \dfrac{1}{1-r_i^2}, i = 1, \ldots, p$ represent the inflation that each regression expression above ideal.



The condition number

It can be measured in terms of the ratio of the largest to the smallest eigenvalue, e.g., the quantity

$$\phi = \frac{\lambda_{\max}}{\lambda_{\min}}$$

which is called the condition number of the correlation matrix. Ratios of eigenvalues, i.e., ratios $\phi_j = \lambda_{\max} / \lambda_j$ are more reliable for diagnosing the impact of a dependency than the eigenvalue λ_j itself.

Consider centered and scaled regressors, $\boldsymbol{X}^*, \boldsymbol{X}^{*T}\boldsymbol{X}^*$ being a correlation matrix and

$$Var\left(\hat{\boldsymbol{\beta}}\right) = \sigma^2\left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)^{-1}, \qquad \sum_{i=1}^{p}\frac{Var\left(\hat{\beta}_i\right)}{\sigma^2} = tr\left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)^{-1} = \sum_{i=1}^{p}\frac{1}{\lambda_i}.$$

The operation given by $\boldsymbol{V}^T\left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)^{-1}\boldsymbol{V} = diag\left(\lambda_1,\ldots,\lambda_p\right)$

$\boldsymbol{V} = \left(\upsilon_1,\ldots,\upsilon_p\right)$ is an orthogonal matrix, is called eigenvalue decomposition of $\boldsymbol{X}^{*T}\boldsymbol{X}^*$.

- If multicollinearity is present, at least one value of j,

$\upsilon_j{}^T\left(\boldsymbol{X}^{*T}\boldsymbol{X}^*\right)^{-1}\upsilon_j\left(\geq \lambda_{\min}\right) \cong 0,$ this implies that for at least one eigenvector υ_j,

$$\sum_{\ell=1}^{p}\upsilon_\ell \boldsymbol{x}_j^* \cong \boldsymbol{0}.$$

VIF in R

- https://www.rdocumentation.org/packages/car/versions/3.0-7/topics/vif
- (install.packages("car")); library(car);

```
set.seed(1082); n = 100 ;
x1 = rgamma(n, shape = 1) ; x2 = rexp(n) ; y = 3+ 2*x1 + x2 + rnorm(n)
vif(lm(y~x1+x2)) ## vifs < 1.1

set.seed(1082); n = 100 ;
x1 = rgamma(n, shape = 1) ; x2 = 2*x1 + rnorm(n, sd = 0.5) ; y = 3+ 2*x1 + x2 + rnorm(n)
vif(lm(y~x1+x2)) ## vifs > 20
```

Alternatives to least squares in cases of multicollinearity

- Ridge regression
- Principal components regression
- Variable selection

Ridge regression

- The ridge regression estimator of the coefficient $\boldsymbol{\beta}$ is found by solving for $\boldsymbol{\beta}_R$ in the system of equations $\left(\boldsymbol{X}^T\boldsymbol{X}+\lambda_R\boldsymbol{I}\right)\boldsymbol{\beta}_R=\boldsymbol{X}^T\boldsymbol{y}$ where $\lambda_R \geq 0$ often referred to as a shrinkage parameter. The solution is given by

$$\hat{\boldsymbol{\beta}}_R=\left(\boldsymbol{X}^T\boldsymbol{X}+\lambda_R\boldsymbol{I}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

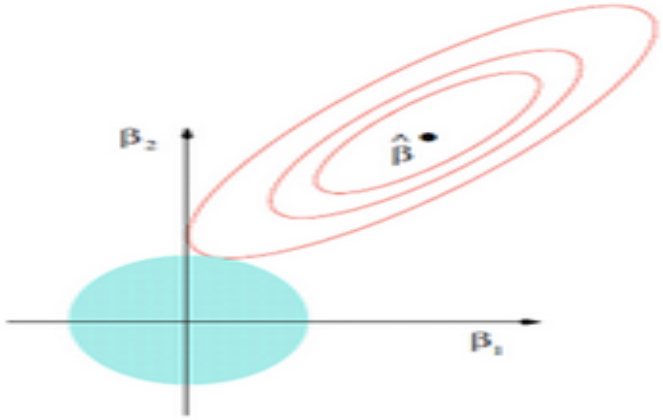
- Consider centered and scaled regressors,

$$\sum_{i=1}^{p}\frac{Var\left(\hat{\beta}_{i,R}\right)}{\sigma^2}=\sum_{i=1}^{p}\frac{\lambda_i}{\left(\lambda_i+\lambda_R\right)^2},\ \lambda_i, i=1,\ldots,p \text{ are the eigenvalue of } \boldsymbol{X}^T\boldsymbol{X}.$$

Ridge Regression:

$$\hat{\beta}^{\text{ridge}} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2,$$

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 \leq t, \qquad \text{(L2 term)}$$



https://rpubs.com/skydome20/R-Note18-Subsets_Shrinkage_Methods

- For example, in the case of the three regressor variables with $\boldsymbol{\lambda}$= (2.985, 0.01, 0.005), least squares estimation gives

$$\sum_{i=1}^{3} \frac{Var\left(\hat{\beta}_i\right)}{\sigma^2} = \sum_{i=1}^{p} \frac{1}{\lambda_i} = 300.335$$

- If rigde regression with say $\lambda_R = 0.1$ is used, the sum of the variances is given by

$$\sum_{i=1}^{3} \frac{\lambda_i}{\left(\lambda_i + \lambda_R\right)^2} \cong 1.6$$

- The procedure would be successful if a λ_R is chosen so that the variance reduction is greater than the bias term given in

$$\sum_{i=1}^{p} \left(E\left(\hat{\beta}_{i,R}\right) - \beta_i \right)^2 = {\lambda_R}^2 \boldsymbol{\beta}^T \left(\boldsymbol{X}^T \boldsymbol{X} + \lambda_R \boldsymbol{I} \right)^{-2} \boldsymbol{\beta}.$$

Shrinkage

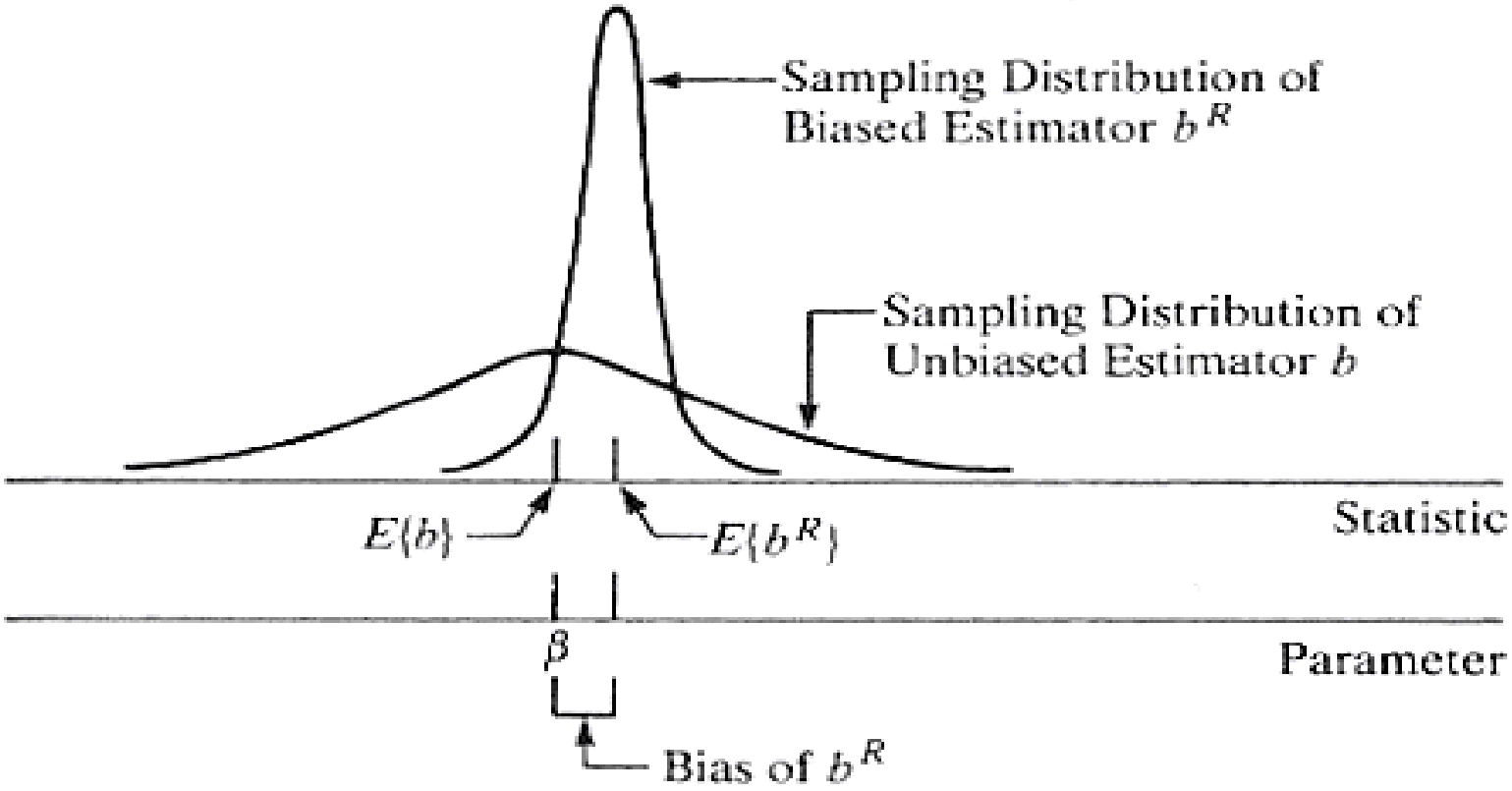
- the similarity to the ordinary least squares solution

  i.e. As $\lambda_R \to \infty$, $\hat{\beta}_R \to \hat{\beta}_{OLS}$.
- In the special case of an orthonormal design matrix,

$$\hat{\beta}_R = \frac{\hat{\beta}_{OLS}}{1+\lambda_R}$$

- This illustrates the essential feature of ridge regression : shrinkage.
- Applying the ridge regression penalty has the effect of shrinking the estimates toward zero - introducing bias but reducing the variance of the estimate

Biased estimator with small variance

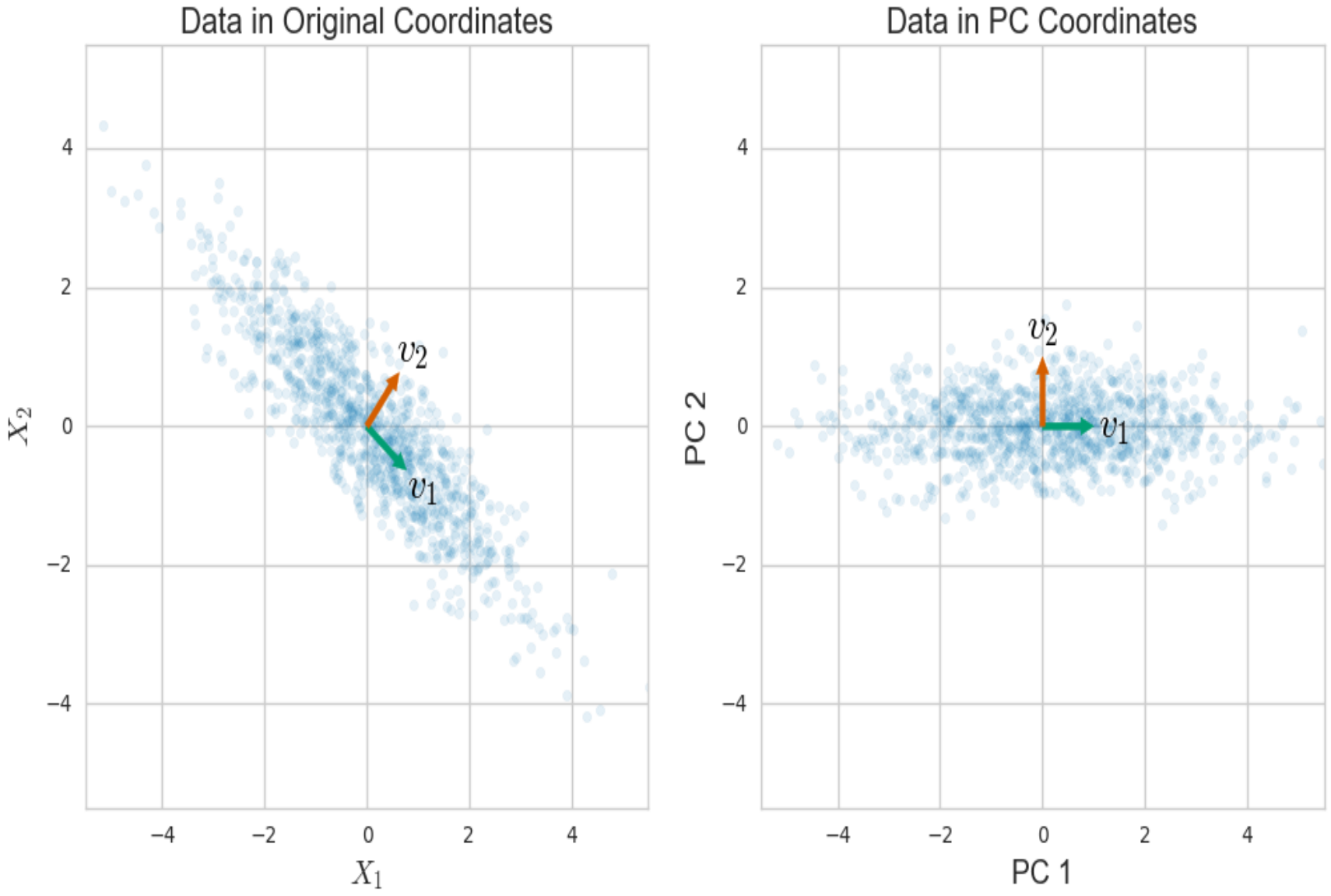


Ridge regression in R

- https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/lm.ridge.html
- install.packages("MASS")
- install.packages("glmnet")

Principal components regression (PCA)

- Consider the original regression model in the form $\boldsymbol{y} = \beta_0 \boldsymbol{1} + \boldsymbol{X}^* \boldsymbol{V}\boldsymbol{V}^T \boldsymbol{\beta} + \boldsymbol{\varepsilon} = \beta_0 \boldsymbol{1} + \boldsymbol{Z}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$. We have $\boldsymbol{Z}^T \boldsymbol{Z} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$ where $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{I}$ and $\boldsymbol{V}$ is an orthogonal matrix.
- $\dfrac{Var\left(\hat{\alpha}_i\right)}{\sigma^2} = \dfrac{1}{\lambda_i}, i = 1, \ldots, p.$


Data in Original Coordinates
Data in PC Coordinates
X_1
X_2
PC 1
PC 2
v_1
v_2
4
2
0
−2
−4

- Suppose we consider the matrix $\boldsymbol{V} = [\nu_1, \ldots, \nu_p]$ of normalized eigenvectors of $\boldsymbol{X}^{*T}\boldsymbol{X}^*$ partitioned into $\boldsymbol{V}=[\boldsymbol{V}_k \vdots \boldsymbol{V}_r]$ and consider Λ to be a diagonal matrix of eigenvalues of $\boldsymbol{X}^{*T}\boldsymbol{X}^*$. We partition Λ as $\Lambda=[\text{diag}(\Lambda_k) \vdots \text{diag}(\Lambda_r)]$.
- An elimination of one (at least one) principal component, that associated with the small eigenvalue, may substantially reduce the total variance in the model and thus produce an appreciably improved prediction equation.
- The difficulty arises in the decision of how many components we should eliminate.

Advantages on principal component regression (PCR)

- There is virtually no limit for the number of predictors.

  i.e. p is allowed to be larger than n
- Correlated or noisy predictors do not undermine the regression fit.
- The PCs carry the maximum amount of variance possible.

Partial least squares (PLS) regression

- Stepwise **select** the orthogonal variables of PCA into the model and coefficients are estimated by **ordinary least squares**
- An alternative to PLS, principal component regression (PCR) i.e. fit using the first *k* principal components from the predictors

  drawback: no guarantee that PCs are associated with the outcome

PCA and PCR in R

- https://www.cnblogs.com/leezx/p/6120302.html
- https://www.jamleecute.com/principal-components-analysis-pca-%E4%B8%BB%E6%88%90%E4%BB%BD%E5%88%86%E6%9E%90/
- https://cran.r-project.org/web/packages/pls/pls.pdf

```
> set.seed(1082); n = 100 ; x1 = rgamma(n, shape = 1) ; x2 = 2*x1 + rnorm(n, sd = 0.25) ; x3 = rexp( n )
> x.input = cbind(x1,x2,x3) ; prcomp( x.input )
Standard deviations (1, .., p=3):
[1] 2.33396740 0.92529449 0.09974213

Rotation (n x k) = (3 x 3):
           PC1          PC2         PC3
x1  0.44365914 -0.007453146  0.89616461
x2  0.89479755 -0.052149124 -0.44341607
x3 -0.05003904 -0.998611496  0.01646738
> biplot( prcomp(x.input), scale = 0) ;  summary( prcomp(x.input))
Importance of components:
                          PC1    PC2     PC3
Standard deviation     2.3340 0.9253 0.09974
Proportion of Variance 0.8628 0.1356 0.00158
Cumulative Proportion  0.8628 0.9984 1.00000
```

R^2 and variable selection

A computing formula for the square of the multiple coefficient coefficient in a $k-$parameter model, J_k, is

$$R^2_{J_K} = \frac{SSR_{J_K}}{SST} = 1 - \frac{SSE_{J_K}}{SST} = 1 - \frac{\hat{\sigma}^2_{J_K}}{\hat{\sigma}^2}.$$