

Coefficient of determination. (R^2)

①

$R^2 = \frac{\text{variation of } Y \text{ explained by the model.}}{\text{Total variation.}}$

$$= \frac{\text{SS Model.}}{\text{SS Total.}} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

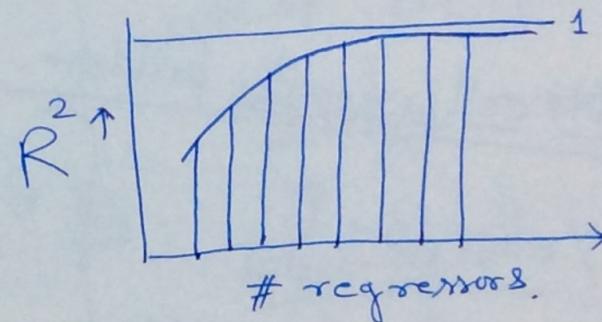
② Larger the value of R^2 , the better the model is

③ $R^2 \in [0, 1]$

$$SST = SS \text{ Model} + SS \text{ Error.}$$

$$\Rightarrow 1 = R^2 + \frac{SS \text{ Error}}{SST.}$$

$$\Rightarrow R^2 = 1 - \frac{SS \text{ Error}}{SS \text{ Total.}}$$



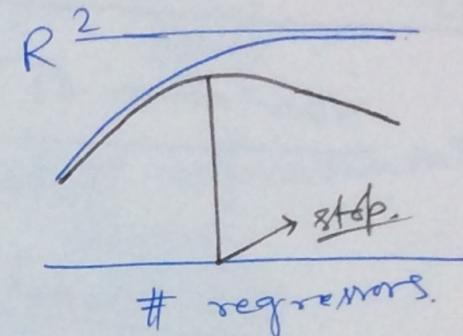
④ If we increase the number of regressor variable, the value of R^2 also increase.

(2)

Adjusted - R^2 (R^2_{adj})

$$R^2_{\text{adj}} = 1 - \frac{\text{SSE}_{\text{true}}/\text{df}(\text{SSE})}{\text{SST}_{\text{true}}/\text{df}(\text{SST}_{\text{true}})}$$

$$= 1 - \frac{\text{SSE}/(n-k-1)}{\text{SST}/(n-1)} < R^2$$



→ Polynomial regression. decide the degree of the model.

Multi collinearity problem.

$$y = x\beta + \epsilon \quad \beta \in \mathbb{R}^{k+1}$$

(ill-condition)

Rank $(x^T x) = k+1$
 Although $|x^T x| \neq 0$ $|x^T x| \approx 0$. \rightarrow problem.

why? then what?

dimensions are similar in nature (why?)

①. columns are very close to zero.
 ②. stratified random sampling is not done properly.

Impact:

(3)

$$\hat{\beta} = (x^T x)^{-1} x^T y = \frac{A_{1j}(x^T x)}{|x^T x|} x^T y.$$

As $|x^T x| \approx 0 \Rightarrow \text{Var}(\hat{\beta}_j)$ may be unbounded. for some j 's

$\Rightarrow |\text{D}(\hat{\beta})| \text{ or } \text{tr}(\text{D}(\hat{\beta}))$ will be too large.

$\Rightarrow \hat{y}_0 = (1 \ x_0)^T \hat{\beta}$ will have large variance.

Consider the expected square error.

$$E[(\hat{\beta} - \beta)^T (\hat{\beta} - \beta)]$$

$$= \sum_{j=0}^K E(\hat{\beta}_j - \beta_j)^2$$

$$= \sum_{j=0}^K \text{Var}(\hat{\beta}_j) \quad \text{tr}(\sigma^2(x^T x)^{-1})$$

$$\hat{\beta} \sim N(\beta, \sigma^2(x^T x)^{-1})$$

\downarrow max eigenvalue.
 $\lambda_0 > \lambda_1 > \lambda_2 \dots \lambda_r > \dots \lambda_K \geq 0$
 nonnegative values.
 arranged in decreasing order. } \rightarrow product going to zero

$$\text{tr}(\sigma^2(x^T x)^{-1}) = \sigma^2 \left(\sum_{j=0}^K \frac{1}{\lambda_j} \right)$$

Note:

(I) $x^T x$ is a symmetric matrix.

(II) $(x^T x)^{-1}$ is also symmetric.

(III) Let the eigenvalues of $(x^T x)$ can be arranged as:

$$\lambda_0 > \lambda_1 > \lambda_2 > \lambda_3 > \dots > 0.$$

$$\begin{aligned} \text{(IV)} \quad \lambda_{\min} &= \min_{\underline{z} \neq 0} \frac{\underline{z}^T (x^T x) \underline{z}}{\underline{z}^T \underline{z}} \geq 0 \\ &\quad \text{symmetric w.r.t. only} \\ &= \min_{\underline{z} \neq 0} \frac{(x \underline{z})^T (x \underline{z})}{\underline{z}^T \underline{z}} \geq 0 \end{aligned}$$

(V) $\text{tr}(x^T x) = \sum_{j=0}^k \lambda_j$

(VI) $\text{tr}((x^T x)^{-1}) = \sum_{j=0}^k \frac{1}{\lambda_j}$

(VII) $|x^T x| = \prod_{j=0}^k \lambda_j$

(VIII) $|(x^T x)^{-1}| = \prod_{j=0}^k \left(\frac{1}{\lambda_j}\right)$

④
SVD / PCA
LA
Stat
Opt

Spectral representation theorem.
Math

We restrict the sum of λ_i such that $\sum_{i=0}^k \frac{1}{\lambda_i}$ remains bounded.

Hence we need to focus on the eigen vector corresponding to $\lambda_0 > \lambda_1 > \dots > \lambda_p$.

Principal component regression.

(5)

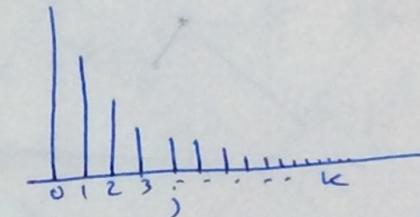
Principal component analysis is dimension reduction technique in multivariate analysis / high dimensional data analysis.

$X^T X$ is a p.s.d matrix / or pd matrix with $|X^T X| \approx 0$.

We can arrange the eigen values in decreasing order.

$$\lambda_0 > \lambda_1 > \lambda_2 \cdots \cdots > \lambda_k \geq 0$$

↓ ↓ ↓ ↓ ↓
 \tilde{u}_0 \tilde{u}_1 \tilde{u}_2 $\cdots \cdots$ \tilde{u}_k .



$$(X^T X) \tilde{u}_j = \lambda_j \tilde{u}_j \quad j = 0, 1, 2, \dots, k.$$

We can represent all these vectors in a matrix form:

$$P = [\tilde{u}_0 \ \tilde{u}_1 \ \cdots \ \tilde{u}_k]$$

\tilde{u}_j 's are orthonormal vectors.

Hence P is an orthogonal matrix.

$T: \mathbb{R}^n \rightarrow \mathbb{R}^n$
there is a matrix A_T

$$T(\underline{v}) = A_T \underline{v}$$

$$sp(\underline{v}) = \{c \underline{v}\} \quad c \in \mathbb{R}$$

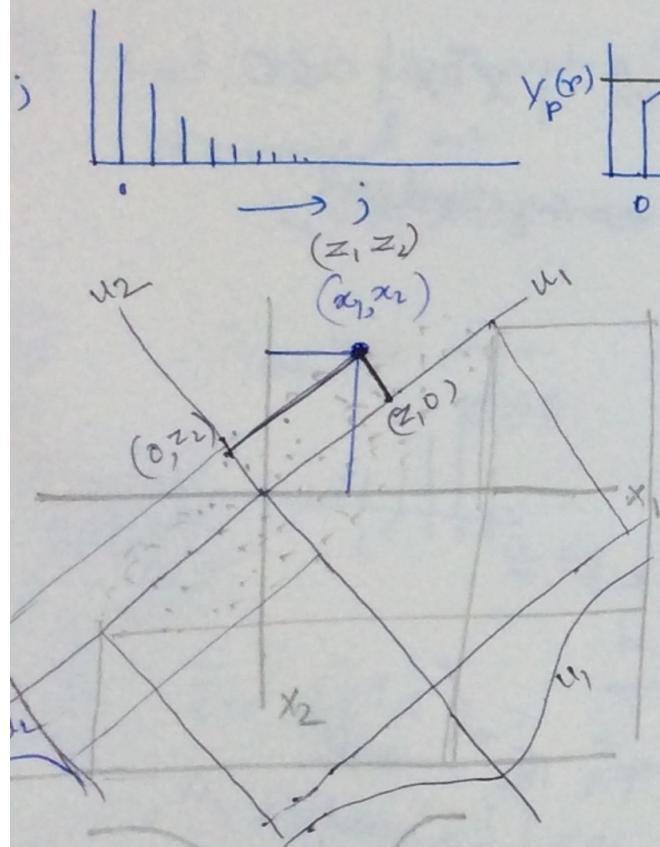
$$A_T \underline{v} = ? \underline{v}$$

(6)

Variation proportion.

$$V_p(r) = \frac{\sum_{i=0}^r \lambda_i}{\sum_{j=0}^k \lambda_j}$$

$$0 \leq r \leq k.$$



No fixed rule.
to choose how many
 (λ_j, v_j) 's to be considered.

$\{v_j\}$ are new
basis elements
which are
orthonormal.

$$\lambda_1 = \lambda_{\max} = \max_{x \neq 0} \frac{x^T A x}{x^T x}$$

$$\lambda_2 = \min_{x \neq 0} \frac{x^T A x}{x^T x}$$

$$x \perp u \text{ when } \lambda_n = \lambda_{\max} u,$$

λ_3 is max in $(\text{int } \{u, u_2\})^\perp$
 λ_3 is max in $(\text{sp } \{u, u_2, u_3\})^\perp$.

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

↓
x₁ axis ↓
x₂ axis.

$$= (x_1, x_2) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} + (x_1, x_2) \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$= \langle z, e_1 \rangle \cdot e_1 + \langle z, e_2 \rangle \cdot e_2$$

$$\underset{(k+1)}{X^T} = P D P^T$$

$$P = [u_0 \ u_1 \ \dots \ u_n] \quad \text{good} \rightarrow$$

④ dimension reduction ⑦.

$$D = \text{diag}(\lambda_0 \ \lambda_1 \ \dots \ \lambda_n) \quad \text{bad} \rightarrow$$

④ can do prediction in old format data.

$$Z = X P$$

$n \times (k+1)$ $n \times (k+1) \ (k+1) \times (k+1)$.

→ new regressions. and we will use $\underline{\underline{\alpha}} = \underline{\underline{Z}} \underline{\underline{\alpha}}$ columns of $\underline{\underline{Z}}$ matrix.

$$Y = X \underline{\underline{\beta}} + \underline{\epsilon}. \quad \underline{\epsilon} \sim N(0, \sigma^2 I_n).$$

$$\Rightarrow Y = X P P^T \underline{\underline{\beta}} + \underline{\epsilon}. \quad P P^T = I.$$

$$\Rightarrow Y = Z \underline{\underline{\alpha}} + \underline{\epsilon}$$

$$\boxed{Y \approx Z_{(r)} \underline{\underline{\alpha}}_{(r)} + \underline{\epsilon}.}$$

↓ reduced model.

$$\hat{\underline{\alpha}}_{(r)} = (Z_{(r)}^T Z_{(r)})^{-1} Z_{(r)}^T Y.$$

One of the ways to address multicollinearity problem.

$$\underline{\underline{\alpha}} = \underline{\underline{Z}} \underline{\underline{\alpha}}$$

we have started with.

$$\underline{\underline{\beta}} \in \mathbb{R}^{k+1}$$

$$\text{but } \hat{\underline{\alpha}}_{(r)} \in \mathbb{R}^{r+1}$$

~~Now~~ Now we have to predict for some new value of $\underline{\underline{Z}}$ which belongs to \mathbb{R}^{n+1}

$$\underline{\underline{\beta}} = I \underline{\underline{\beta}}.$$

$$\Rightarrow \underline{\underline{\beta}} = P P^T \underline{\underline{\beta}}.$$

$$\Rightarrow \underline{\underline{\beta}} = P \underline{\underline{\alpha}}$$

$$\Rightarrow \hat{\underline{\beta}}_{pc} = P \begin{pmatrix} \underline{\alpha}_{(r)} \\ \vdots \\ \underline{\alpha}_r \\ \vdots \\ \underline{\alpha}_k \end{pmatrix}$$

$$\hat{Y} = \hat{\underline{\beta}}_{pc}^T \underline{\underline{x}}_{\text{new}}$$

If A and B both are invertible. then.

$$\underline{(AB)^{-1} = B^{-1}A^{-1}}$$

$$\frac{X}{m \times (k+1)} \text{ is not invertible}$$

$$\underline{X^T X}$$

$$\underline{(k+1) \times (k+1)}$$

Shrinkage method / Regularization.

Multicollinearity problem. $|X^T X| \approx 0$.

$\hat{\beta}_{LS}$ is an unbiased estimator of β .

But the components of $\hat{\beta}$ may have very large variation.

or equivalently $\|\hat{\beta}_{LS}\|_2^2 = \sum_{i=0}^k \hat{\beta}_i^2$ is large. (L_2 norm)

* We want to have an estimate of $\hat{\beta}$ such that.
 L_1 norm. $\|\hat{\beta}\|_1 = \sum_{i=0}^k |\hat{\beta}_i|$ or L_2 norm $\|\hat{\beta}\|_2 = \left(\sum_{i=0}^k \hat{\beta}_i^2 \right)^{1/2}$

or $\|\hat{\beta}\|_2^2$ are bounded.

$\|\hat{\beta}_0\|_1 < C_1$
LASSO estimate.

$\|\hat{\beta}\|_2 < C_2$
Ridge estimate.

(9).

Let \underline{x} be a random variable with

$$E(|x|^{\infty}) < \infty \quad \text{then} \quad \text{where } k \leq \infty$$

$$E(|x|^k) < \infty$$

$$\|x\|_p = \left(\sum_{i=1}^k |x_i|^p \right)^{1/p} = \|\underline{x}\|_p.$$

$$\|\hat{\beta}\|_2^2 \leq c. \quad \text{hyper sphere.}$$

We want to obtain the value $\alpha \beta$ such that $\perp S$ condition

$$S(\beta) = (\underline{y} - \alpha \beta)^T (\underline{y} - \alpha \beta)$$
 is minimized. w.r.t. $\sum_{i=0}^k \beta_i^2 \leq c.$

we Lagrange Multiplier

$$\hat{\beta}_R = \arg \min_{\beta} (S(\beta) + \lambda (\beta^T \beta - c))$$

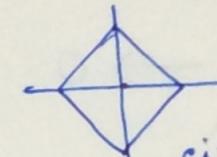
$$\hat{\beta}_R = (X^T X + \lambda I)^{-1} X^T Y$$

(Ridge estimate $\alpha \beta$)
is a function of c .



L1 norm.

$$\|\beta\|_1 < c.$$



circle
under L2 norm:

$$x^2 + y^2 \leq r^2 \rightarrow \text{circle.}$$

$$x^2 + y^2 + z^2 \leq r^2 \rightarrow \text{sphere.}$$

$$x^2 + y^2 + z^2 + w^2 \leq r^2 \rightarrow \text{hypercube.}$$

$$\hat{\beta}_R = (x^T x + \lambda I)^{-1} x^T y.$$

Case 1 $\lambda \rightarrow \infty$ then $\hat{\beta}_R \rightarrow 0$.

$\lambda \rightarrow 0$ then $\hat{\beta}_R \rightarrow \hat{\beta}_{LS}$.

① Is $\hat{\beta}_R$ an unbiased estimator of β ? NO.

$$\begin{aligned}\hat{\beta}_R &= (x^T x + \lambda I)^{-1} x^T y \\ &= (x^T x + \lambda I)^{-1} (x^T x) \underbrace{(x^T x)^{-1} x^T y}_{\hat{\beta}_{LS}} \\ &= [(x^T x + \lambda I)^{-1} (x^T x)] \hat{\beta}_{LS}.\end{aligned}$$

$$E(\hat{\beta}_R) = \underbrace{[(x^T x + \lambda I)^{-1} (x^T x)]}_{\text{Not an identity matrix.}} \beta.$$

Bias is introduced to reduce the variance,

① Does $\hat{\beta}_R$ has less dispersion compared to $\hat{\beta}_{LS}$? YES.

Show $D(\hat{\beta}_{LS}) - D(\hat{\beta}_R)$ is p.s.d.

$$= \sigma^2 (S^{-1} - WS S^{-1} SW)$$

$$= \sigma^2 (S^{-1} - WSW)$$

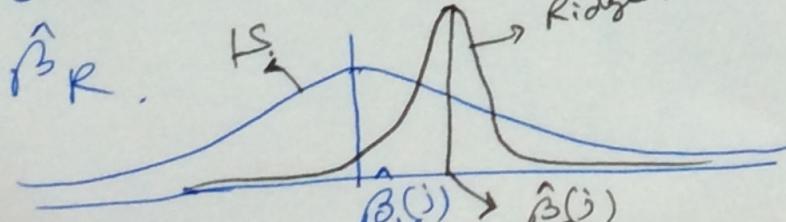
$$= \sigma^2 W (W^{-1} S^{-1} W^{-1} - S) W$$

$$= \sigma^2 W ((S + \lambda I) S^{-1} (S + \lambda I) - S) W$$

$$= \sigma^2 W [2\lambda I + \lambda^2 S^{-1}] W$$

If λ is large.
 $\lambda^2 > 2\lambda$. If p.s.d. ~~not~~

Show Dispersion of $\hat{\beta}_{LS}$ is more than the dispersion of $\hat{\beta}_R$.



$$W = (X^T X + \lambda I)^{-1}$$

$$S = (X^T X)$$

$$W^{-1} = (S + \lambda I)$$

$$D(\hat{\beta}_{LS}) = S^{-1} \sigma^2$$

$$D(\hat{\beta}_R) = D(W S \hat{\beta}_{LS})$$

$$\hat{\beta}_{LS} = \underbrace{(X^T X)}_A^{-1} \underbrace{X^T Y}_L$$

$$D(\hat{\beta}_{LS}) = A D(Y) A^T$$

$$= (X^T X)^{-1} X^T (\underbrace{I \sigma^2}_L) X (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1} (X^T X) (X^T X)^{-1}$$

$$= \sigma^2 (X^T X)^{-1}$$

$$= \underline{\sigma^2 S^{-1}}$$

