Coefficient of determination. (R^2)

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

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

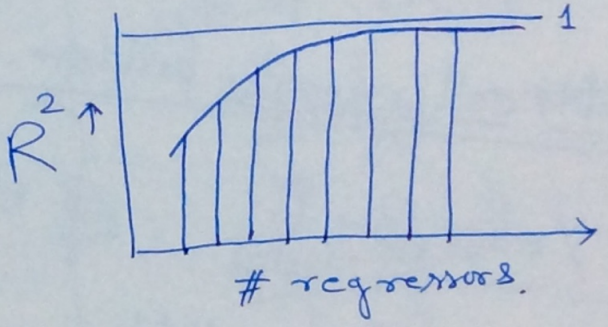
(*) Larger the value of R^2, the better the model is

(*) $R^2 \in [0, 1]$

$$SST = SS\ Model + SS\ Error.$$

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

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

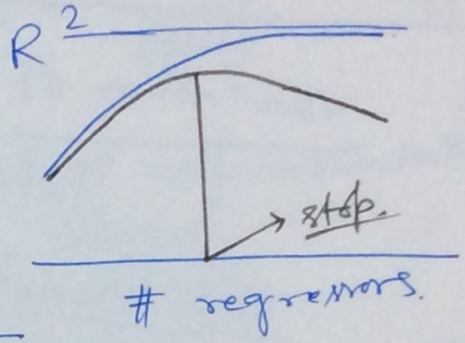


(*) If we increase the numbers of regressor variable, the value of R^2 also increase.

Adjusted-R^2 $\left(R^2_{adj}\right)$

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

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



→ Polynomial regression. decide the degree of the model.

Multi collinearity problem.

$Y = X\beta + \epsilon$ $\quad \beta \in \mathbb{R}^{k+1}$

Rank $(X^TX) = k+1$

Although $|X^TX| \neq 0$ $\quad |X^TX| \approx 0$. (ill conditioned) → problem.

(Reason.)

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:

$$\hat{\beta} = (X^TX)^{-1}X^TY = \frac{Adj(X^TX)}{|X^TX|}X^TY.$$

As $|X^TX| \approx 0 \Rightarrow Var(\hat{\beta}_j)$ may be unbounded. for some j's

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

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

Consider the expected square error:

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

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

$$= \sum_{j=0}^{k} Var(\hat{\beta}_j) = tr\left(\sigma^2(X^TX)^{-1}\right) = \sigma^2\left(\sum_{j=0}^{k}\frac{1}{\lambda_j}\right)$$

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

↓ max eigen vm.

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

non negative values arranged in decreasing order. } → product going to zero

Note:

(i) X^TX is a symmetric matrix.

(ii) $(X^TX)^{-1}$ is also symmetric.

(iii) Let the eigenvalues of (X^TX) can be arranged as.

$$\lambda_0 > \lambda_1 > \lambda_2 > \lambda_3 \cdots \geq 0.$$

(iv) $$\lambda_{min} = \min_{\underline{z} \neq \underline{0}} \frac{\underline{z}^T (X^TX)\underline{z}}{\underline{z}^T\underline{z}} \geq 0$$

Symmetric matrix only

$$= \min_{\underline{z} \neq \underline{0}} \frac{(X\underline{z})^T (X\underline{z})}{\underline{z}^T\underline{z}} \geq 0$$

(v) $$tr(X^TX) = \sum_{j=0}^{k} \lambda_j$$

(vi) $$tr\left((X^TX)^{-1}\right) = \sum_{j=0}^{k} 1/\lambda_j$$

(vii) $$|X^TX| = \prod_{j=0}^{k} \lambda_j$$

(viii) $$\left|(X^TX)^{-1}\right| = \prod_{j=0}^{k} (1/\lambda_j).$$

SVD / PCA

LA → Start

01

Spectral representation theorem.

Math

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

Hence we need to focus on the eigenvectors corresponding to $\lambda_0 > \lambda_1 \cdots \lambda_r$.

Principal component regression.

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

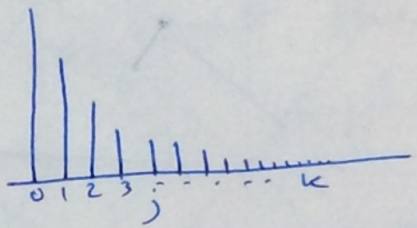
X^TX is a p.s.d matrix / or pd matrix with $|X^TX| \approx 0$.

We can arrange the eigen values in decreasing order.

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

$$\downarrow \quad \downarrow \quad \downarrow \qquad \downarrow$$

$$\underset{\sim}{u}_0 \quad \underset{\sim}{u}_1 \quad \underset{\sim}{u}_2 \cdots \underset{\sim}{u}_k.$$



$$(X^TX)\,\underset{\sim}{u}_j = \lambda_j \underset{\sim}{u}_j \qquad j = 0, 1, 2, \cdots k.$$

We can represent all these vectors in a matrix form.

$$P = [u_0\ u_1 \cdots u_k]$$

u_j's are orthonormal vectors.
Hence P is an orthogonal matrix.

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

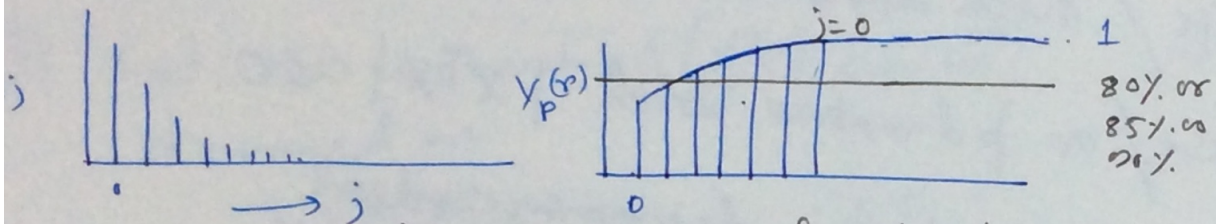
$$T(\underset{\sim}{v}) = A_T \underset{\sim}{v}.$$

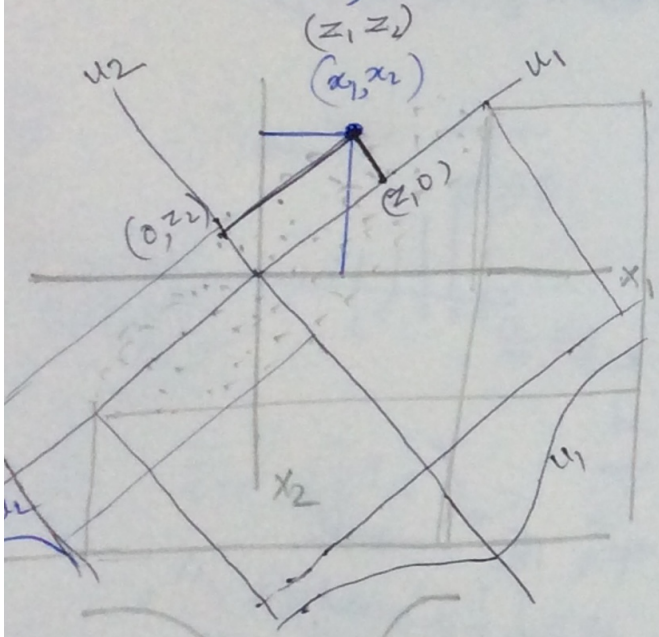
$$sp(\underset{\sim}{v}) = \{c\underset{\sim}{v}\} \qquad c \in \mathbb{R}$$

$$A_T \underset{\sim}{v} = \lambda \underset{\sim}{v}.$$

Variation proportion.

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





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

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

$$\lambda_1 = \lambda_{max} = \max_{x \neq 0} \frac{x^T A x}{x^T x} \qquad \text{Spectral decomposition theorem.}$$

$$\lambda_2 = \max_{\substack{x \neq 0 \\ x \perp u \\ \text{wh.}}} \frac{x^T A x}{x^T x} \qquad A u_1 = \lambda_{max} u_1$$

λ_3 is max in $(\text{sp}\{u_1, u_2\})^{\perp}$

λ_3 is max in $(\text{sp}\{u_1, u_2, u_3\})^{\perp}$.

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

x_1 axis $\qquad$ x_2 axis

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

$$= \langle \underset{\sim}{x}, e_1 \rangle \cdot e_1 + \langle \underset{\sim}{x}, e_2 \rangle \cdot e_2$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \langle \underset{\sim}{x}, \underset{\sim}{u}_1 \rangle \underset{\sim}{u}_1 + \langle \underset{\sim}{x}, \underset{\sim}{u}_2 \rangle \underset{\sim}{u}_2$$

$$= z_1 \underset{\sim}{u}_1 + z_2 \underset{\sim}{u}_2$$

$X^T X = P D P^T$

$P = [u_0\ u_1\ \cdots\ u_k]$ good →

$D = \text{diag}(\lambda_0\ \lambda_1\ \cdots\ \lambda_k)$

good →
- ⊛ dimension reduction
- ⊛ can do prediction in old form of data.

bad →
- ⊛ interpretation of regressors are lost
- ⊛ $\underset{\sim}{Y}$ has no role in this dimension reduction.

$$\underset{n\times(k+1)}{Z} = \underset{n\times(k+1)}{X}\ \underset{(k+1)\times(k+1)}{P}$$

→ new regressors. and we will use 'r' ($r+1$) colums of Z matrix.

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

$$PP^T = I.$$

Let $P^T\underset{\sim}{\beta} = \underset{\sim}{\alpha}$.

$$\Rightarrow Y = XPP^T\underset{\sim}{\beta} + \underset{\sim}{\epsilon}.$$

$$\Rightarrow Y = Z\underset{\sim}{\alpha} + \underset{\sim}{\epsilon}$$

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

↓ reduced model.

$$\alpha_{(r)} = \begin{pmatrix}\alpha_0\\ \alpha_1\\ \vdots\\ \alpha_r\end{pmatrix} \qquad \underset{\sim}{\alpha} = \begin{pmatrix}\alpha_0\\ \vdots\\ \alpha_k\end{pmatrix}$$

$r < k$.

$$\hat{\underset{\sim}{\alpha}}_{(r)} = \left(Z_{(r)}^T Z_{(r)}\right)^{-1} Z_{(r)}^T \underset{\sim}{Y}.$$

One of the ways to address multicollinearity problem.

we have started with.

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

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

~~Note~~ Now we have to predict for some new value of X which belongs to $\mathbb{R}^{k+1}$

$$\underset{\sim}{\beta} = I\underset{\sim}{\beta}.$$

$$\Rightarrow \underset{\sim}{\beta} = PP^T\underset{\sim}{\beta}.$$

$$\Rightarrow \underset{\sim}{\beta} = P\underset{\sim}{\alpha}$$

$$\Rightarrow \hat{\underset{\sim}{\beta}}_{PC} = P\begin{pmatrix}\alpha_{(r)}\\ 0\\ \vdots\\ 0\end{pmatrix}$$

$\hat{y} = \hat{\beta}_{PC}^T x$ → new

If A and B both are invertable. then.

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

$\underset{n\times(k+1)}{X}$ is not invertable

X^TX

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

Shrinkage method / Regularization.

Multicollinearity or problem. $|X^TX| \approx 0$.

$\hat{\beta}_{LS}$ 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 β 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.

$\|\hat{\beta}_*\|_1 < C_1$ or $\|\hat{\beta}\|_2 < C_2$

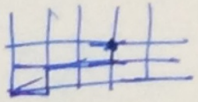
LASSO estimate. Ridge estimate.

Let $\underline{x}$ be a random variable with.

$E(|x|^r) < \infty$ then.

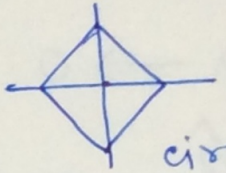
$E(|x|^k) < \infty$ where $0 \le k \le r$

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



L_1 norm.

$\|\beta\|_1 < C$.



circle under L_1 norm.

$\|\hat{\beta}\|_2^2 \le C$. hyper sphere.

$x^2+y^2 \le r^2 \rightarrow$ circle.
$x^2+y^2+z^2 \le r^2 \rightarrow$ sphere.
$x^2+y^2+z^2+w^2 \le r^2 \rightarrow$ hypersphere.

We want to obtain the value of β such that LS condition

$S(\beta) = (\underline{Y} - X\underline{\beta})^T (\underline{Y} - X\underline{\beta})$ is minimized. w.r.t. $\sum_{i=0}^{k} \beta_i^2 \le C$.

Use Lagrange Multiplier.

$$\hat{\beta}_R = \arg\min_{\beta} \left(S(\underline{\beta}) + \lambda (\underline{\beta}^T \underline{\beta} - C) \right).$$

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

(Ridge estimate of β is a function of C.)

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

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

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

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

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

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

$$= \left[(X^T X + \lambda I)^{-1}(X^T X)\right] \hat{\beta}_{LS}.$$

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

Bias is introduced to reduce the variance.

(*) 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.

$$= \sigma^2 \left(S^{-1} - WSS^{-1}SW \right)$$

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

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

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

$$= \sigma^2 W \underbrace{\left[2\lambda I + \lambda^2 S^{-1} \right]}_{\downarrow \text{ p.s.d}} W.$$

if λ is large.

$\lambda^2 > 2\lambda$. If p.s.d.

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

$S = (X^TX)$

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

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

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

$\hat{\beta}_{LS} = \underbrace{(X^TX)^{-1}X^T}_{A} Y.$

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

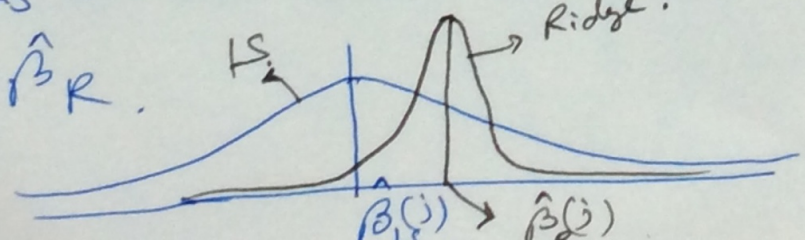
$= (X^TX)^{-1}X^T (I\sigma^2) X (X^TX)^{-1}$

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

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

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

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



Bias

$$E(\hat{\beta}_R) - \beta$$

$$= WS\beta - \beta$$

$$= (WS - I)\beta$$

$$= -\lambda W \beta$$

$$MSE(\hat{\beta}_R)$$

$$= \text{var}(\hat{\beta}_R) + \text{Bias}^2$$

$$= \text{tr}\left(D(\hat{\beta}_R)\right) + \beta^T (WS-I)^T (WS-I)\beta$$

(total variance; Bias^2)

$$= \sigma^2 \text{tr}(WSW) + \lambda^2 \beta^T W^2 \beta$$

$$= \sigma^2 \text{tr}\left[P(D+\lambda I)^{-1} D (D+\lambda I)^{-1} P^T\right] + \lambda^2 \beta^T (PDP^T + \lambda I)^{-2} \beta$$

$$= \sigma^2 \sum_{i=0}^{k} \frac{\lambda_i}{(\lambda_i+\lambda)^2} + \lambda^2 \sum_{i=0}^{k} \frac{\beta_i^2}{(\lambda_i+\lambda)^2}$$

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

$\Rightarrow I = WS + \lambda W$.

$\Rightarrow -\lambda W = WS - I$.

$S = X^T X$

$= PDP^T$

$D = (\lambda_0, \lambda_1, \dots, \lambda_k)$ ↓ diagonal matrix.

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

$= (PDP^T + \lambda PP^T)^{-1}$

$= (P(D+\lambda I)P^T)^{-1}$

$= P(D+\lambda I)^{-1} P^T$.

$MSE = \text{Var} + \text{Bias}^2$

$T(x)$ is an unbiased estimator of θ.

$T_2(x)$ is a biased estimator of θ.

does it imply

$MSE_\theta(T_2(x)) > MSE_\theta(T_1(x))$??

Not always possible.

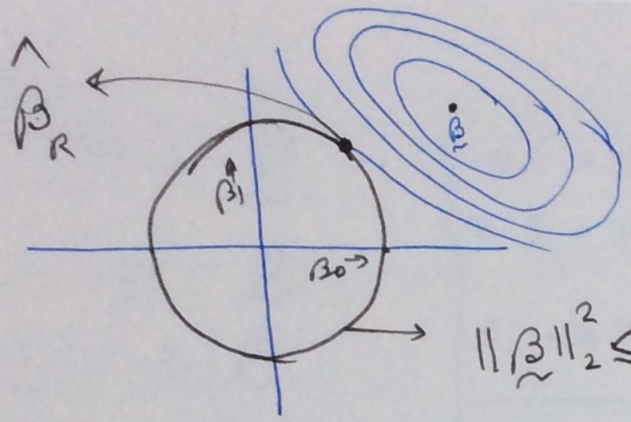
$T_1(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$ unbiased

$T_2(x) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ biased.

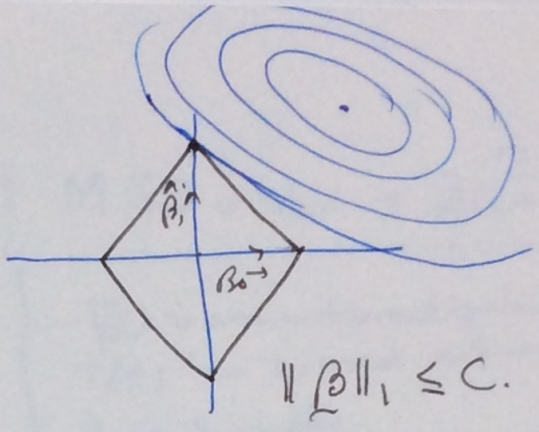
$\text{tr}\ \sigma^2$.

$\text{tr}(A) = \sum \text{eigenvalues}$

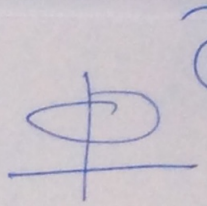
$A^T = A \Rightarrow A = P D P^T$ → diagonal with eigenvalues.



Ridge regression.



LASSO.



Lasso can help for dimension reduction.
Ridge can not help dimension reduction.

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}.$$

$$\{x \mid x \in \mathbb{R}^n, \ f(x) = k\} = \text{contour of } f(x) = k.$$