Math B7800

2/1/18, Thur

Chatterjee

Advanced Topics in Stat. (Prof. Shirshendu Chatterjee)

Textbook: Applied Multivariate Analysis

Applied Linear Statistical model by Kuther, etc.

Topics

Multiple Linear Regression and associated inference.

Logistic regression, ~~principle~~ principal component analysis, clustering and classification, some design of experiment.

HW: Biweekly HW will be posted in Blackboard.

Solution of some of them will be posted.

Exams: Two Midterms (40%) ← Tentatively on the first class of March and April.

One project (20%) ← Due by the last class.

Final Exam (40%)

Makeup Exam (strict policy)

Formula sheets (3 pages for Midterms, 4 pages for Final and calculators are allowed)

Ch.11 from J & W.

Multiple linear regression

Y ⟷ dependent variable (response)

$Z_1, \dots, Z_K$ ⟷ K independent variable (predictors)

Assume linear relation : $Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_K Z_K + \varepsilon$ (error)

Goal: statistical inference about $\beta_0, \dots, \beta_K$

Prediction of response based on predictor variables.

We observe the response and predictors for n observations.

This is called **the training data**.

check

$$
\begin{aligned}
Y_1 &= \beta_0 + \beta_1 Z_{11} + \beta_2 Z_{12} + \cdots + \beta_K Z_{1K} + \varepsilon_1 \\
Y_2 &= \beta_0 + \beta_1 Z_{21} + \beta_2 Z_{22} + \cdots + \beta_K Z_{2K} + \varepsilon_2 \\
&\vdots \\
Y_n &= \beta_0 + \beta_1 Z_{n1} + \beta_2 Z_{n2} + \cdots + \beta_K Z_{nK} + \varepsilon_n
\end{aligned}
$$

$$
\underset{\sim}{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad Z = \begin{bmatrix} 1 & Z_{11} & \cdots & Z_{1K} \\ \vdots & \vdots & & \vdots \\ 1 & Z_{n1} & \cdots & Z_{nK} \end{bmatrix}, \quad \underset{\sim}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \quad \underset{\sim}{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_K \end{pmatrix}
$$

$$
\underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times(k+1)}{Z}\underset{(k+1)\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}}
$$

Assumption on errors

Errors have mean 0 and they are uncorrelated.

Remember that the random part in the model is the errors, $\underset{\sim}{\varepsilon}$.

$\mathbb{E}(\varepsilon_i) = 0$ for all i.

$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$ for all $i \neq j$.

$\mathbb{E}(\varepsilon_i \varepsilon_j) - \mathbb{E}(\varepsilon_i)\mathbb{E}(\varepsilon_j)$

$\mathrm{var}(\varepsilon_i) = \mathrm{cov}(\varepsilon_i, \varepsilon_i) = \sigma^2$ (homoskedasticity assumption)

In this case, we can use **least square method** to estimate $\underset{\sim}{\beta}$.

Minimize $\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 Z_{i1} - \cdots - \beta_K Z_{iK})^2$ — (*)

$\underset{\sim}{\varepsilon} = \underset{\sim}{Y} - Z\underset{\sim}{\beta}$

$\mathbb{E}(Y_i) = \beta_0 + \beta_1 Z_{i1} + \cdots + \beta_K Z_{iK}$.

The minimizer of (*) is called the LS estimator of $\underset{\sim}{\beta}$

Notation $\underset{\sim}{\hat{\beta}}_{LS}$.

the sum of squared error of differences.

$= \sum_{i=1}^{n} (Y_i - Z_{i,*}\underset{\sim}{\beta})^2$, $Z_{i,*}$ is the i^{th} row.

$= (\underset{\sim}{Y} - Z\underset{\sim}{\beta})'(\underset{\sim}{Y} - Z\underset{\sim}{\beta}) = \underset{\sim}{\varepsilon}'\underset{\sim}{\varepsilon}$

$\frac{\partial}{\partial \underset{\sim}{\beta}} (\underset{\sim}{Y} - Z\underset{\sim}{\beta})'(\underset{\sim}{Y} - Z\underset{\sim}{\beta}) = 0$.

$2Z'(\underset{\sim}{Y} - Z\underset{\sim}{\beta}) = 0 \Rightarrow Z'\underset{\sim}{Y} = Z'Z\underset{\sim}{\beta}$.

$\underset{\sim}{\hat{\beta}}_{LS} = (Z'Z)^{-1}Z'\underset{\sim}{Y}$.

$\mathbb{E}(\underset{\sim}{\varepsilon}) + Z\underset{\sim}{\beta} = \underset{\sim}{Y}$

$Z'Z\underset{\sim}{\beta} = Z'\underset{\sim}{Y}$

$\underset{\sim}{\hat{\beta}} = (Z'Z)^{-1}Z'\underset{\sim}{Y}$

If $\text{rank}(Z) = k+1$, then $\hat{\beta}_{LS} = (Z'Z)^{-1}Z'Y$.

$\hat{Y}$ = predicted value of Y

$= Z\hat{\beta} = Z(Z'Z)^{-1}Z'Y = P_Z Y$, where

P_Z is the projection matrix on the column space of Z. ($P_Z = Z(Z'Z)^{-1}Z'$)

Recall that projection matrices are **idempotent**,

i.e., $P_Z^2 = P_Z$. Also, $I - P_Z$ is idempotent. $(I-P_Z)^2 = I - P_Z$.

Matrix A is Idempotent iff $A^2 = A$

e.g. $A = \begin{bmatrix} 4 & -1 \\ 12 & -3 \end{bmatrix}$, $A^2 = \begin{bmatrix} 4 & -1 \\ 12 & -3 \end{bmatrix}$

Also, $P_Z(I - P_Z) = 0$. $\therefore P_Z \perp (I - P_Z)$

Thus, column space of P_Z and $I - P_Z$ are orthogonal.

$E(\hat{\beta}_{LS}) = E((Z'Z)^{-1}Z'Y)$

$= (Z'Z)^{-1}Z'E(Y)$

$= (Z'Z)^{-1}Z'Z\beta = \beta$ ($Y = Z\beta + \varepsilon$)

$\text{cov}(\hat{\beta}) = \text{cov}((Z'Z)^{-1}Z'Y) = (Z'Z)^{-1}Z'\,\text{cov}(Y)\,Z(Z'Z)^{-1}$ ($\text{cov}(Y) = \text{cov}(\varepsilon)$)

$\text{cov}(AY) = A\,\text{cov}(Y)A'$

$= (Z'Z)^{-1}Z'(\sigma^2 I)Z(Z'Z)^{-1}$

$= \sigma^2 (Z'Z)^{-1}$

- **Unbiased estimator for σ^2**

The estimated error $\hat{\varepsilon} = Y - \hat{Y}$

$= Y - P_Z Y = (I - P_Z)Y = \hat{\varepsilon}$

So, $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}'\hat{\varepsilon} = Y'(I-P_Z)^2 Y$ ← $(I-P_Z)^2 = (I-P_Z)$

$= Y'(I - P_Z)Y$

$E(\hat{\varepsilon}'\hat{\varepsilon}) = E[Y'(I-P_Z)Y]$

★ $x'Ax = tr(x'Ax) = tr(Axx')$ if A is a square symmetric $(k\times k)$ matrix and x is $(k\times 1)$ vector

$= E[tr(Y'(I-P_Z)Y)]$

$= E[tr((I-P_Z)YY')]$

$= tr(E((I-P_Z)YY'))$

$= tr[(I-P_Z)E(YY')]$ ($E(YY') = var(Y) + E(Y)E(Y')$, $E(Y) = Z\hat{\beta}$, $E(Y') = \hat{\beta}'Z'$, $E(Y)E(Y') = Z\hat{\beta}\hat{\beta}'Z'$)

$= tr[(I-P_Z)(\sigma^2 I + Z\hat{\beta}\hat{\beta}'Z')]$ ($Z\hat{\beta} = P_Z Y$)

$= \sigma^2 tr[(I-P_Z) + 0]$ By $(I-P_Z)Z\hat{\beta}\hat{\beta}'Z' = 0$ $\because (I-P_Z)P_Z = 0$

$= \sigma^2 tr(I - P_Z) = \sigma^2 \text{rank}(I - P_Z) = \sigma^2(n-k-1)$

Since, $(I-P_Z)P_Z = 0$,

$\text{rank}(P_Z) + \text{rank}(I - P_Z) = n$

Also, $\text{rank}(Z) = \text{rank}(P_Z) = $ # columns of Z assuming that Z has full column rank.

$= k+1$

So, $\text{rank}(I - P_Z) = n - (k+1)$.

So, an unbiased estimator of σ^2 is

$$\frac{\hat{\underset{\sim}{\varepsilon}}'\hat{\underset{\sim}{\varepsilon}}}{n-k-1} = \frac{\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y}}{n-k-1} = \hat{\sigma}^2$$

Under the "mean 0 uncorrelated" assumption on the errors, we found unbiased estimator of $\underset{\sim}{\beta}$ and σ^2.

Math B7800 2/6/18, Tue

$$S(\underset{\sim}{\beta}) = (\underset{\sim}{Y}-Z\underset{\sim}{\beta})'(\underset{\sim}{Y}-Z\underset{\sim}{\beta}) = \sum_{i=1}^{n}(Y_i - Z_i)$$

$$\underset{\sim}{\beta} = (\beta_0, \beta_1, \dots, \beta_r) \qquad = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 z_{i1} - \cdots - \beta_r z_{ir})^2$$

$$\begin{bmatrix} \frac{\partial S}{\partial \beta_0} \\ \vdots \\ \frac{\partial S}{\partial \beta_r} \end{bmatrix} = \frac{\partial S}{\partial \underset{\sim}{\beta}}$$

Last time we looked at the least square estimate of $\underset{\sim}{\beta}$

$$\hat{\underset{\sim}{\beta}}_{LS} = (Z'Z)^{-1}Z'\underset{\sim}{Y}$$

Model

$$\underset{n\times 1}{\underset{\sim}{Y}} = \underset{n\times(r+1)}{Z}\underset{(r+1)\times 1}{\underset{\sim}{\beta}} + \underset{n\times 1}{\underset{\sim}{\varepsilon}}$$

◉ Decomposition of sum of squares.

Total sum of Square $= \sum_{i=1}^{n} Y_i^2$

Total SS $\underset{\sim}{Y}'\underset{\sim}{Y}$ $\underset{\sim}{Y}'(I-P_{\underset{\sim}{1}})\underset{\sim}{Y}$ $= (\underset{\sim}{Y}'\underset{\sim}{Y})$. $\underset{\sim}{Y}'P_{\underset{\sim}{1}}\underset{\sim}{Y}$

$$\sum_{i=1}^{n} Y_i^2 = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 + n\bar{Y}^2$$

total SS (corrected) or total SS about the mean.

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}Y_i^2 + \underbrace{\sum_{i=1}^{n}\bar{Y}^2}_{n\bar{Y}^2} - \underbrace{2\sum_{i=1}^{n}Y_i\bar{Y}}_{2n\bar{Y}^2}$$

$$= \sum_{i=1}^{n}Y_i^2 - n\bar{Y}^2 \qquad (n\bar{Y})$$

$$\bar{Y} = \frac{1}{n}\underset{\sim}{1}_n'\underset{\sim}{Y}, \quad \underset{\sim}{1}_n = \underset{n \text{ many}}{\begin{bmatrix}1\\ \vdots \\ 1\end{bmatrix}}$$

in terms of Matrix?

$$\begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} - \begin{pmatrix} \bar{y} \\ \vdots \\ \bar{y} \end{pmatrix} = y - 1_n \bar{y} = y - \frac{1}{n} 1_n 1_n' y$$

($\frac{1}{n} 1_n' y$)

y − $\bar{y}$

$$= \left(I - \frac{1}{n} 1_n 1_n'\right) y$$

$$= (I - P_{1_n}) y$$

$$P_{1_n} = 1_n (1_n' 1_n)^{-1} 1_n' = \frac{1}{n} 1_n 1_n'$$

(“n”)

So, $(I - P_{1_n})^2 = (I - P_{1_n})$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}' \begin{pmatrix} y_1-\bar{y} \\ \vdots \\ y_n-\bar{y} \end{pmatrix}$$

Total SS about the mean

$$= y'(I - P_{1_n})^2 y = y'(I - P_{1_n}) y$$

idempotent

Total SS about mean $= y'(I - P_{1_n}) y$

$\sum_{i=1}^{n} (y_i - \bar{y})^2$. ? $= y'(I - P_Z) y + y'(P_Z - P_{1_n}) y$

Residual SS — Regression SS

$\phi \hat{\varepsilon}'\hat{\varepsilon}$ = (remains: if bigger ⇒ prediction is failed.) — (Model is good or not: if it is bigger, good)

$$y'(I - P_Z) y = y'(I - P_Z)^2 y$$

“residual SS” $= y'(I - P_Z)(I - P_Z) y$

$(I - P_Z) y = y - P_Z y = y - \hat{y}$ (because $P_Z y = Z\hat{\beta} = \hat{y}$)

$= \hat{\varepsilon}$

$= \hat{\varepsilon}'\hat{\varepsilon} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

residual SS

$= \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 z_{i1} - \cdots - \hat{\beta}_r z_{ir})^2$

“Regression SS”

☆ $y'(P_Z - P_{1_n}) y = ?$

$P_Z 1_n = 1_n$, then

$P_Z P_{1_n} = P_Z \frac{1}{n} 1_n 1_n' = \frac{1}{n} P_Z 1_n 1_n'$

P_{1_n} $= \frac{1}{n} 1_n 1_n' = P_{1_n}$

In general, they are not equal. But here they are equal

i.e. $P_Z P_{1_n} \neq P_{1_n} P_Z$ “Not Commute”

So, $(P_Z - P_{1_n})^2 = P_Z^2 + P_{1_n}^2 - P_Z P_{1_n} - P_{1_n} P_Z$

$(P_Z - P_{1_n})$ $= P_Z + P_{1_n} - P_{1_n} - P_{1_n}$

“Regression SS” $= (P_Z - P_{1_n})$ “idempotent”

$$y'(P_Z - P_{1_n}) y = y'(P_Z - P_{1_n})(P_Z - P_{1_n}) y = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

$(P_Z - P_{1_n}) y = \hat{y} - \bar{y} = \begin{pmatrix} \hat{y}_1 - \bar{y} \\ \vdots \\ \hat{y}_n - \bar{y} \end{pmatrix}$

$P_Z y - P_{1_n} y$

$\hat{y}$ $\bar{y}$

Also observe $(I-P_Z)(P_Z-P_1)=0$.

Also, $(I-P_1)P_1=0$

So, the decomposition of $\underset{\sim}{Y}$ into $(I-P_1)\underset{\sim}{Y}$ and $P_1\underset{\sim}{Y}$ is an orthogonal decomposition.

Also, $(I-P_1)\underset{\sim}{Y}=(I-P_Z)\underset{\sim}{Y}+(P_Z-P_1)\underset{\sim}{Y}$

This is also orthogonal decomposition.

$\underset{\sim}{Y}=Z\underset{\sim}{\beta}+\underset{\sim}{\varepsilon}$ ... (1)

$\underset{\sim}{\varepsilon}=\underset{\sim}{Y}-P_1\underset{\sim}{Y}=\underset{\sim}{Y}-\bar{\underset{\sim}{Y}}$

$\underset{\sim}{Y}-P_Z\underset{\sim}{Y}=(\underset{\sim}{Y}-\hat{\underset{\sim}{Y}})=\hat{\underset{\sim}{\varepsilon}}$, $P_Z\underset{\sim}{Y}-P_1\underset{\sim}{Y}=(\hat{\underset{\sim}{Y}}-\bar{\underset{\sim}{Y}})=\underset{\sim}{\varepsilon}_{reg}$

$\underset{\sim}{\varepsilon}=\hat{\underset{\sim}{\varepsilon}}+\underset{\sim}{\varepsilon}_{reg}$

$P_Z\underset{\sim}{Y}=\hat{\underset{\sim}{Y}}$, $\underset{\sim}{Y}-\hat{\underset{\sim}{Y}}=\hat{\underset{\sim}{\varepsilon}}$ ... (2)

- A measure of goodness of the regression is

$$R^2=\frac{\text{Reg. SS}}{\text{Tot. SS about Mean}}=\frac{\underset{\sim}{Y}'(P_Z-P_1)\underset{\sim}{Y}}{\underset{\sim}{Y}'(I-P_1)\underset{\sim}{Y}}$$

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

$$=1-\frac{\text{Res. SS}}{\text{Total SS about mean}}$$

$1-R^2=\frac{\text{Residual SS}}{\text{Tot. SS about Mean}}=\frac{\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y}}{\underset{\sim}{Y}'(I-P_1)\underset{\sim}{Y}}$

$P_Z\underset{\sim}{Y}-\bar{\underset{\sim}{Y}}=\hat{\underset{\sim}{Y}}-\bar{\underset{\sim}{Y}}$, $\underset{\sim}{Y}-\hat{\underset{\sim}{Y}}=\hat{\underset{\sim}{\varepsilon}}$

$\therefore R^2$ lies between 0 and 1.

R^2 not good fit; R^2 good fit

If R^2 is close to 1, then the regression model is a good fit

(predictor has no information for $\underset{\sim}{Y}$)

If R^2 is close to 0, then the model is that the predictors can not predict the response well.

$R^2=1$ only when $\hat{\varepsilon}_i=0$ for all i.

This means $y_i=\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 z_{i1}+\cdots+\hat{\beta}_r z_{ir}$.

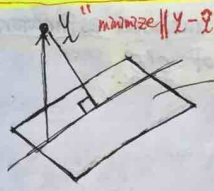
$R^2=0$, when $\hat{\beta}_0=\bar{y}$, $\hat{\beta}_1=\cdots=\hat{\beta}_r=0$.

$R=+\sqrt{R^2}$ is called Multiple correlation coefficient.

- Geometric Point of view for the LS method.

The LS problem:

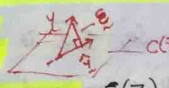
minimize $\|\underset{\sim}{Y}-Z\underset{\sim}{\beta}\|^2=(\underset{\sim}{Y}-Z\underset{\sim}{\beta})'(\underset{\sim}{Y}-Z\underset{\sim}{\beta})$ with respect to $\underset{\sim}{\beta}$



" $\{Z\underset{\sim}{b}: \underset{\sim}{b}$ is $(r+1)\times 1$ vector $\}$.

So, the minimization problem is equivalent to orthogonal projection of $\underset{\sim}{Y}$ onto $C(Z)$

So, if $Z\underset{\sim}{b}$ is the closest to $\underset{\sim}{Y}$ among all points of $C(Z)$, then $\underset{\sim}{Y}-Z\underset{\sim}{b}$ is orthogonal to $C(Z)$



So, $(\underset{\sim}{Y}-Z\underset{\sim}{b})$ is orthogonal to $Z\underset{\sim}{c}$ for any $\underset{\sim}{c}\in C(Z)$

So, $\underset{\sim}{c}'Z'(\underset{\sim}{Y}-Z\underset{\sim}{b})=0$ for all $\underset{\sim}{c}$ $\quad(\underset{\sim}{Y}-Z\underset{\sim}{b})\perp Z\underset{\sim}{c}$

So, $Z'(\underset{\sim}{Y}-Z\underset{\sim}{b})=0 \Rightarrow \underset{\sim}{b}=(Z'Z)^{-1}Z'\underset{\sim}{Y}$

Hence, $Z\underset{\sim}{b}=P_Z\underset{\sim}{Y}$ by $Z\underset{\sim}{b}=\underbrace{Z(Z'Z)^{-1}Z'}_{P_Z}\underset{\sim}{Y}$

• Best Linear Unbiased Estimator (BLUE)

For any vector $\underset{\sim}{c}$ (column vector), suppose we want to estimate $\underset{\sim}{c}'\underset{\sim}{\beta}$

An estimator will be called linear if it has the form $\underset{\sim}{a}'\underset{\sim}{Y}$ for some $\underset{\sim}{a}$

Thm (Gauss)

$\underset{\sim}{c}'\hat{\underset{\sim}{\beta}}_{LS}$ is the best among all linear unbiased estimators of $\underset{\sim}{c}'\underset{\sim}{\beta}$.

Best in the sense "smallest variance"

$\mathbb{E}[\underset{\sim}{c}'\hat{\underset{\sim}{\beta}}] = \underset{\sim}{c}'\mathbb{E}[\hat{\underset{\sim}{\beta}}] = \underset{\sim}{c}'\underset{\sim}{\beta}$.

So, $\underset{\sim}{c}'\hat{\underset{\sim}{\beta}}$ is unbiased for $\underset{\sim}{c}'\underset{\sim}{\beta}$. ($\hat{\beta}$ 가 β 의 unbiased estimator 이면 $a_*'Z\hat{\beta}$ 가 $a_*'Z\beta$ 의 unbiased estimator 이므로 $a_*'Z = c'$)

$\underset{\sim}{c}'\hat{\underset{\sim}{\beta}} = \underset{\sim}{c}'(Z'Z)^{-1}Z'\underset{\sim}{Y} = \underset{\sim}{a}_*'\underset{\sim}{Y}$, where $\underset{\sim}{a}_* = Z(Z'Z)^{-1}\underset{\sim}{c}$

(linear) (linear)

So, $\underset{\sim}{c}'\hat{\underset{\sim}{\beta}}$ is also linear estimator.

If $\underset{\sim}{a}'\underset{\sim}{Y}$ is unbiased for $\underset{\sim}{c}'\underset{\sim}{\beta}$, then ⇔ $\mathbb{E}[\underset{\sim}{a}'\underset{\sim}{Y}] = \mathbb{E}[\underset{\sim}{a}'Z\beta + \underset{\sim}{a}'\underset{\sim}{\varepsilon}]$

↳ $\mathbb{E}[\underset{\sim}{a}'\underset{\sim}{Y}] = \underset{\sim}{c}'\underset{\sim}{\beta}$ for all $\underset{\sim}{\beta}$ ($\underset{\sim}{c}'\underset{\sim}{\beta}$ = $\underset{\sim}{a}'Z\underset{\sim}{\beta}$)

So, $\underset{\sim}{a}'Z\underset{\sim}{\beta} = \underset{\sim}{c}'\underset{\sim}{\beta}$ for all $\underset{\sim}{\beta}$ So, $\underset{\sim}{a}'Z = \underset{\sim}{c}'$

(by ⊛) ($\mathbb{E}[Y]$ 또는 $(c'-a'Z)\beta = 0, \forall\beta$)

$\mathrm{Var}(\underset{\sim}{a}'\underset{\sim}{Y}) = \underset{\sim}{a}'\mathrm{var}(Y)\,\underset{\sim}{a} = \sigma^2\underset{\sim}{a}'\underset{\sim}{a}$.

$\mathrm{var}(\underset{\sim}{c}'\hat{\underset{\sim}{\beta}}) = \mathrm{var}(a_*'\underset{\sim}{Y}) = \sigma^2 a_*'a_*$

$\underset{\sim}{a}'\underset{\sim}{a} = (\underset{\sim}{a} - a_* + a_*)'(\underset{\sim}{a} - a_* + a_*)$

$= (\underset{\sim}{a} - a_*)'(\underset{\sim}{a} - a_*) + a_*'a_* + \underbrace{2(\underset{\sim}{a} - a_*)'a_*}_{=0} \geq a_*'a_*$.

(by) $(\underset{\sim}{a} - a_*)'Z(Z'Z)^{-1}\underset{\sim}{c} = 0$ since $(\underset{\sim}{a} - a_*)'Z = \underset{\sim}{c}' - \underset{\sim}{c}' = 0$ ($a'Z - a_*'Z$)

Math B7800 2/8/18, Thur

$\underset{\sim}{Y} = (I - P_{\underset{\sim}{1}})\underset{\sim}{Y} + P_{\underset{\sim}{1}}\underset{\sim}{Y}$ "orthogonal decomposition"

BLUE (continued)

$a_* = Z(Z'Z)^{-1}\underset{\sim}{c}$

$\underset{\sim}{a}$ satisfies $\underset{\sim}{a}'Z = \underset{\sim}{c}'$

Need to check $(\underset{\sim}{a} - \underset{\sim}{a}_*)'\underset{\sim}{a}_* = 0$.

$\underset{\sim}{c}'\hat{\underset{\sim}{\beta}} = \underset{\sim}{c}'(Z'Z)^{-1}Z'\underset{\sim}{Y} = \underset{\sim}{a}_*'\underset{\sim}{Y} = \underset{\sim}{a}_*'Z\hat{\underset{\sim}{\beta}}$

so, $\underset{\sim}{c}' = \underset{\sim}{a}_*'Z$.

$(\underset{\sim}{a} - \underset{\sim}{a}_*)'Z(Z'Z)^{-1}\underset{\sim}{c} = 0$, since $(\underset{\sim}{a} - \underset{\sim}{a}_*)'Z = \underset{\sim}{c}' - \underset{\sim}{c}' = 0$ ($\underset{\sim}{a}'Z - \underset{\sim}{a}_*'Z$)

• Inference about Regression

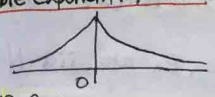
So far we didn't assume any distribution for the error ε_i's.

We only assumed $\mathbb{E}(\varepsilon_i) = 0$ and $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0 & \text{if } i \neq j \\ \sigma^2 & \text{if } i = j \end{cases}$.

Now, we need to assume distribution of the error.

1. Suppose the errors are iid Double exponential dist.

with pdf $f(x) = \frac{1}{2}e^{-|x|}$, $x \in \mathbb{R}$ ($\sigma^2 = 1$)



Suppose $\varepsilon_1, \ldots, \varepsilon_n$ are iid with pdf $f(x)$.

Q: Find MLE of $\underset{\sim}{\beta}, \sigma^2$.

dist. of $\underset{\sim}{Y}$?

Model $\underset{\sim}{Y} = Z\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$.

Suppose $Y_1, Y_2, \ldots, Y_n$ are independent.

$\mathbb{E}(Y_i) = Z_{i,*}\underset{\sim}{\beta} = \beta_0 + \beta_1 Z_{i1} + \cdots + \beta_r Z_{ir}$

↓

dist. of Y_i ?

PDF of Y_i ?

Pdf of $\varepsilon_i = Y_i - Z_{i,*}\beta$ is $f(x)$ ie Double exponential dist.

$$f(Y_i) = \frac{1}{2} e^{-|Y_i - \underbrace{Z_{i,*}\beta}_{\hat{Y}}|}$$

So, joint pdf of $Y_1, \dots, Y_n$.

$$f(Y_1, \dots, Y_n) = \prod_{i=1}^{n} f(Y_i) = \frac{1}{2^n} e^{-\sum_{i=1}^{n} |Y_i - Z_{i*}\beta|} = L(\beta)$$

"Not continuous!" "likelihood function"

Maximizing $L(\beta)$ is equivalent to

Minimizing $\sum_{i=1}^{n} |Y_i - \beta_0 - \beta_1 Z_{i1} - \cdots - \beta_r Z_{ir}|$ w.r.t β.

The estimator that minimizes the last term is called LAD (Least absolute deviation) estimator of β

Note Suppose the errors are normally distributed.

So, $\varepsilon \sim N_n(0, \sigma^2 I_n)$

Normal Dist $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

So, $\varepsilon_1, \dots, \varepsilon_n$ are iid $N(0, \sigma^2)$

Q: Find MLE of β, σ^2.

Distribution of Y_i is $N(Z_{i*}\beta, \sigma^2)$

pdf of Y_i $f(Y_i) = \frac{1}{\sqrt{2\pi}\cdot\sigma} e^{-\frac{1}{2\sigma^2}(Y_i - Z_{i*}\beta)^2}$

Joint pdf of $Y_1, \dots, Y_n$

$$f(Y_1, \dots, Y_n) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\sum_{i=1}^{n} \frac{1}{2\sigma^2}(Y_i - Z_{i*}\beta)^2} = L(\beta, \sigma^2)$$

$$\left(\frac{(2\pi)^{-n/2}}{(\sigma^2)^{n/2}}\right)$$

$$\ell(\beta, \sigma^2) = \log L(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - Z_{i*}\beta)^2$$

$$\frac{\partial}{\partial \beta}\ell(\beta, \sigma^2) = -\frac{1}{2\sigma^2}\frac{\partial}{\partial \beta}\sum_{i=1}^{n}(Y_i - Z_{i*}\beta)^2 = 0$$

$\Rightarrow 0$

minimizer

So, $\hat{\beta} = \hat{\beta}_{LS} = (Z'Z)^{-1}Z'Y$. MLE of β?

$$\frac{\partial}{\partial \sigma^2}\ell(\beta, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - Z_{i*}\beta)^2 = 0.$$

So, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - Z_{i*}\hat{\beta})^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon}$

Hence, $\hat{\beta}, \hat{\sigma}^2$ are the MLE.

$\hat{\sigma}^2$ is MLE, but not unbiased for σ^2.

The unbiased estimator for σ^2 is $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-r-1} = s^2$.

$\hat{\beta}$ is unbiased for β.

$$\hat{\beta} = (Z'Z)^{-1}Z'Y \sim N_{r+1}\left(\underbrace{(Z'Z)^{-1}Z'Z}_{I}\beta, (Z'Z)^{-1}Z'(\sigma^2 I)Z(Z'Z)^{-1}\right)$$

$$Y \sim N_n(Z\beta, \sigma^2 I)$$

$$= N_{r+1}\left(\beta, \sigma^2(Z'Z)^{-1}\right)$$

Fact $\hat{\beta}$ and $\hat{\sigma}^2$ (or s^2) are independent. ?

We will show $\hat{\beta}$ and $\hat{\varepsilon}$ are independent.

$\hat{\varepsilon} = (I - P_Z)Y$, $\hat{\beta} = (Z'Z)^{-1}Z'Y$.

$\mathrm{cov}(\hat{\varepsilon}, \hat{\beta}) = (I - P_Z)(\sigma^2 I)\, Z(Z'Z)^{-1} = 0$.

since $(I - P_Z)Z = 0$.

$\mathrm{cov}(AY, BZ) = A\,\mathrm{cov}(Y, Z)B'$

So, $\hat{\varepsilon}$, $\hat{\beta}$ are independent.

Since $\hat{\sigma}^2$ and s^2 are both function of $\hat{\varepsilon}$, they are independent with $\hat{\beta}$.

Q: Distribution of $\hat{\sigma}^2$ and s^2?

$R^2 = \dfrac{Y'(P_Z - P_1)Y}{Y'(I - P_1)Y}$

$1 - R^2 = \dfrac{Y'(I - P_Z)Y}{Y'(I - P_1)Y}$

Fact: $\hat{\varepsilon}'\hat{\varepsilon} \sim \chi^2_{n-r-1}$.

- **Linear Algebra Digression**

Suppose A is symmetric real matrix $(n \times n)$.

If $\lambda_1, \lambda_2, \cdots, \lambda_n$ are the eigenvalues of A, then eigenvalues of A^2 are $\lambda_1^2, \cdots, \lambda_n^2$.

If $A = U\Lambda U'$ is the spectral decomposition.

,i.e, U is orthogonal and $\Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$,

then $A^2 = U\Lambda U' U\Lambda U' = U\Lambda^2 U'$, $\Lambda^2 = \begin{pmatrix} \lambda_1^2 & & 0 \\ & \ddots & \\ 0 & & \lambda_n^2 \end{pmatrix}$

This shows that $\lambda_1^2, \cdots, \lambda_n^2$ are the eigen values of A.

If A is idempotent, then each eigenvalue is either 0 or 1.

because if $A = U\Lambda U'$, then $A = A^2$

$\Rightarrow U\Lambda U' = U\Lambda^2 U'$.

$\Rightarrow \Lambda = \Lambda^2$

$\Rightarrow \lambda_i = \lambda_i^2 \quad \forall i \in \mathbb{N}$.

So, λ_i = either 0 or 1.

So, $A = U \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} U'$, $U = \underset{n\times k}{[U_1} \;\; \underset{n\times(n-k)}{U_2]}$

$= \underset{n\times k}{U_1}\,\underset{k\times n}{U_1'}$, then $\mathrm{rank}(A) = k$.

$\hat{\varepsilon}'\hat{\varepsilon} = Y'(I - P_Z)Y$. Since $(I - P_Z)$ is idempotent and $\mathrm{rank}(I - P_Z) = n - r - 1$

(idempotent ?)

So, $I - P_Z = UU'$, where U is $n \times (n-r-1)$.

and $U'U = I$.

So, $Y'(I - P_Z)Y = Y'UU'Y$ ($\hat{\varepsilon}'\hat{\varepsilon}$; $Y'U$, $U'Y$)

So, $\dfrac{Y'Y}{\sigma^2} \sim \chi^2_{n-r-1}$

$= \dfrac{Y'(I - P_Z)Y}{\sigma^2} = \dfrac{\hat{\varepsilon}'\hat{\varepsilon}}{\sigma^2}$.

Now, see that $U'Y \sim N(U'Z\beta, \sigma^2 I)$ ($U'Y$; $U'Z\beta = 0$)

Also, $UU'Z = (I - P_Z)Z = 0$.

So, $U'UU'Z = 0$

So, $U'Z = 0$.

$U'Y \sim N_{n-r-1}(0, \sigma^2 I)$

$U'Y$

Math B7800 2/13/18, Tue

Model: $\underset{\sim}{Y} = Z\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$, where $\underset{\sim}{\varepsilon} \sim N_n(\underset{\sim}{0}, I_n\sigma^2)$.

Last time $\frac{\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y}}{\sigma^2} \sim \chi^2_{n-r-1}$ (n−r−1: d.f.; r: rank of $(I-P_Z)$)

Also, $\hat{\beta}$ and $\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y}$ are independent.

Goal: Obtain Confidence Region for $\underset{\sim}{\beta}$. Then find Scheffe CI and Bonferroni CI for β_i.

Recall: $\hat{\underset{\sim}{\beta}} = (Z'Z)^{-1}Z'\underset{\sim}{Y} \sim N_{r+1}(\underset{\sim}{\beta}, \sigma^2(Z'Z)^{-1})$

Recall: For symmetric positive definite matrix A, we can define $A^{1/2}$.

Q: Confidence Region for β.

So, $(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}) \sim N_{r+1}(\underset{\sim}{0}, \sigma^2(Z'Z)^{-1})$

$\frac{1}{\sigma}(Z'Z)^{1/2}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}) \sim N_{r+1}(\underset{\sim}{0}, I_{r+1})$

So, $\underbrace{\left[\frac{1}{\sigma}(Z'Z)^{1/2}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})\right]'}_{\underset{\sim}{U}'} \underbrace{\frac{1}{\sigma}(Z'Z)^{1/2}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})}_{\underset{\sim}{U}} = \underset{\sim}{U}'\underset{\sim}{U} \sim \chi^2_{r+1}$.

($\frac{1}{\sigma^2}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})'(Z'Z)(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta}) \sim \chi^2_{r+1}$)

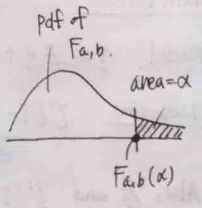
So, using independent of $\hat{\underset{\sim}{\beta}}$ and $\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y}$.

$$F = \frac{\frac{1}{r+1}\frac{1}{\sigma^2}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})'(Z'Z)(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})}{\frac{1}{n-r-1}\frac{1}{\sigma^2}\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y}} \sim F_{r+1,\,n-r-1}$$

($\frac{1}{n-r-1}\frac{1}{\sigma^2}\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y} = S^2$)

$$= \frac{1}{r+1}\frac{1}{S^2}(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})'(Z'Z)(\hat{\underset{\sim}{\beta}} - \underset{\sim}{\beta})$$

Notation : $F_{r+1,n-r-1}(\alpha)$. $F_{a,b}(\alpha)$



So, $\mathbb{P}\left(F \le F_{r+1,n-r-1}(\alpha)\right) = 1-\alpha$.

Thus, $\left\{\underset{\sim}{\beta} : (\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta})'(Z'Z)(\hat{\underset{\sim}{\beta}}-\underset{\sim}{\beta}) \le (r+1)s^2 F_{r+1,n-r-1}(\alpha)\right\}$

is a $100(1-\alpha)\%$ Confidence Region for $\underset{\sim}{\beta}$.

Next topic is CI for β_i "only i, one interval."

Distribution of $\hat{\beta}_i$?

$\hat{\beta}_i \sim N\left(\beta_i, \sigma^2(Z'Z)^{-1}_{i+1,i+1}\right)$ for $i=0,1,\dots,r$.

Note that $(Z'Z)^{-1}_{i+1,i+1}$ refers to the $(i+1),(i+1)$ entry of $(Z'Z)^{-1}$.

Now, $\dfrac{\hat{\beta}_i-\beta_i}{\sqrt{\sigma^2(Z'Z)^{-1}_{i+1,i+1}}} \sim N(0,1)$.

σ^2 unknown. Since σ^2 is unknown, we cannot use it for a CI of β_i. We replace σ^2 by s^2.

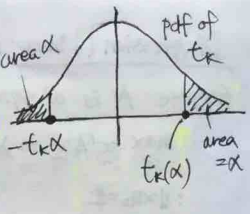
$\dfrac{\hat{\beta}_i-\beta_i}{\sqrt{s^2(Z'Z)^{-1}_{i+1,i+1}}} \sim t_{n-r-1}$.

$= \dfrac{(\hat{\beta}_i-\beta_i)\sqrt{\sigma^2(Z'Z)^{-1}_{i+1,i+1}}}{\sqrt{s^2/\sigma^2}} \sim N(0,1)$

$\dfrac{s^2}{\sigma^2} = \dfrac{1}{n-r-1}\dfrac{Y'(I-P_Z)Y}{\sigma^2} \sim \chi^2_{n-r-1}$.

So, a $100(1-\alpha)\%$ CI for β_i is

$\hat{\beta}_i \pm \sqrt{s^2(Z'Z)^{-1}_{i+1,i+1}}\, t_{n-r-1}\left(\frac{\alpha}{2}\right)$



$\hat{\beta}_i \pm \sqrt{\widehat{var}(\hat{\beta}_i)}\, t_{n-r-1}\left(\frac{\alpha}{2}\right)$

$var(\hat{\beta}_i) = \sigma^2(Z'Z)^{-1}_{i+1,i+1}$. ($\sigma^2$ unknown)

$\widehat{var}(\hat{\beta}_i) = s^2(Z'Z)^{-1}_{i+1,i+1}$. (estimator)

- Simultaneous CI for β_i, $i=0,1,\dots,K$.

Two standard methods

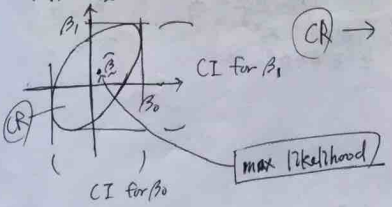
(1) Bonferroni , (2) Scheffe's method.

(1) Bonferroni simultaneous CI for $\beta_0,\beta_1,\dots,\beta_K$ with confidence $1-\alpha$ is

$\hat{\beta}_i \pm \sqrt{\widehat{var}(\hat{\beta}_i)}\, t_{n-r-1}\left(\frac{\alpha}{2(K+1)}\right)$

(2) Scheffe's method, we need to project the $(1-\alpha)100\%$ CR onto different axes.

Suppose $\underset{\sim}{\beta}$ has 2 components.



Diagression (Linear Algebra)

Suppose A is a symmetric matrix.

$$\max_{\underset{\sim}{x}:\|x\|_2=1} \underset{\sim}{x}'A\underset{\sim}{x} = \lambda_1(A) \quad \text{(largest eigenvalue of } A\text{)}$$

$$= \max_{\underset{\sim}{x}:\|x\|_2\le 1} \underset{\sim}{x}'A\underset{\sim}{x} = \lambda_1(A).$$

Suppose A is symmetric, B is symmetric and pd (positive definite).

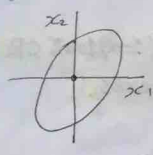
$$\max_{\underset{\sim}{x}:\underset{\sim}{x}'B\underset{\sim}{x}\le 1} (\underset{\sim}{x}'A\underset{\sim}{x}) = \max_{\underset{\sim}{y}:\|\underset{\sim}{y}\|_2\le 1} \underset{\sim}{y}'B^{-1/2}AB^{-1/2}\underset{\sim}{y}$$

Define $\underset{\sim}{y} = B^{1/2}\underset{\sim}{x}$

$$= \lambda_1(B^{-1/2}AB^{-1/2})$$

Next, for two matrices C and D for which CD, DC are both ~~defini~~ defined, then all nonzero eigenvalues of CD and DC are same.

~~#eigenvalues of CD = #eigenvalues of DC~~



$\underset{\sim}{x}'B\underset{\sim}{x} \le 1$

Finding the projection onto x_1-axes.

$$\Leftrightarrow \max_{\underset{\sim}{x}:\underset{\sim}{x}'B\underset{\sim}{x}\le 1} x_1^2$$

$$\max_{\underset{\sim}{x}:\underset{\sim}{x}'B\underset{\sim}{x}\le 1} x_1^2$$

Finding the projection onto x_1-axis.

$$\Leftrightarrow \max_{\underset{\sim}{x}:\underset{\sim}{x}'B\underset{\sim}{x}\le 1} x_1^2$$

$$= \max_{\underset{\sim}{x}:\underset{\sim}{x}'B\underset{\sim}{x}\le 1} \underset{\sim}{x}'A\underset{\sim}{x}, \text{ where } A = \begin{pmatrix}1&0\\0&0\end{pmatrix} = \begin{pmatrix}1\\0\end{pmatrix}(1\ 0)$$

$$= \underset{\sim}{e}_1\underset{\sim}{e}_1', \text{ where } \underset{\sim}{e}_k \text{ is } k\text{th unit vector.}$$

$$= \lambda_1(B^{-1/2}AB^{-1/2})$$

$$= \lambda_1(B^{-1/2}\underset{\sim}{e}_1\underset{\sim}{e}_1'B^{-1/2})$$

$$= \lambda_1(\underset{\sim}{e}_1'B^{-1}\underset{\sim}{e}_1) = \underset{\sim}{e}_1'B^{-1}\underset{\sim}{e}_1 = (B^{-1})_{1,1}$$

($\lambda_1(\underset{\sim}{e}_1'B^{-1}\underset{\sim}{e}_1)$: $\underset{\sim}{e}_1'B^{-1}\underset{\sim}{e}_1$ is a scalar.)

$B^{-1} = \begin{pmatrix}\square & \\ & \end{pmatrix}$.

So, we see that

$$\max_{\underset{\sim}{x}:\underset{\sim}{x}'B\underset{\sim}{x}\le 1} x_i^2 = (B^{-1})_{i,i}.$$

Recall 100(1−α)% CR for $\underset{\sim}{\beta}$ was

$$(\underset{\sim}{\beta}-\hat{\underset{\sim}{\beta}})'(\underset{\sim}{z}'\underset{\sim}{z})(\underset{\sim}{\beta}-\hat{\underset{\sim}{\beta}}) \le (r+1)s^2F_{r+1,n-r-1}(\alpha).$$

Write in the form $\underset{\sim}{x}'B\underset{\sim}{x} \le 1$, where

$$\underset{\sim}{x} = \underset{\sim}{\beta}-\hat{\underset{\sim}{\beta}}, \quad B = (\underset{\sim}{z}'\underset{\sim}{z})\frac{1}{(r+1)s^2F_{r+1,n-r-1}(\alpha)}.$$

$\max |\beta_i - \hat{\beta}_i|^2 = (B^{-1})_{i+1,i+1}$, $i = 0, 1, \dots, r$

$\underset{\sim}{\beta}$: β is in CR.

$= (r+1) s^2 F_{r+1, n-r-1}(\alpha) (Z'Z)^{-1}_{i+1,i+1}$.

So, the Scheffe simultaneous CI for $\beta_0, \dots, \beta_r$ is

$\hat{\beta}_i \pm \sqrt{(r+1) s^2 F_{r+1,n-r-1}(\alpha) (Z'Z)^{-1}_{i+1,i+1}}$.

$= \hat{\beta}_i \pm \sqrt{(r+1) F_{r+1,n-r-1}(\alpha) \widehat{var}(\hat{\beta}_i)}$

Scheffe simultaneous CI for all linear combinations of the components of $\underset{\sim}{\beta}$, namely $\underset{\sim}{a}'\underset{\sim}{\beta}$ for all vectors $\underset{\sim}{a}$ is given by

$\underset{\sim}{a}'\hat{\underset{\sim}{\beta}} \pm \sqrt{(r+1) F_{r+1,n-r-1}(\alpha) \widehat{var}(\underset{\sim}{a}'\hat{\underset{\sim}{\beta}})}$, where

$\widehat{var}(\underset{\sim}{a}'\hat{\underset{\sim}{\beta}}) = \underset{\sim}{a}' \widehat{var}(\hat{\underset{\sim}{\beta}}) \underset{\sim}{a}$

$= s^2 \underset{\sim}{a}' (Z'Z)^{-1} \underset{\sim}{a}$

$P_Z = Z(Z'Z)^{-1}Z'$

$P_{Z_1} = Z_1(Z_1'Z_1)^{-1}Z_1'$

$Z = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$

$Z_1 = \begin{bmatrix} 1 & 0 \\ 3 & 0 \end{bmatrix}$

Math B7800

2/15/18, Thur

March 06, Tue – Exam 01. (3 pages formular sheet)

• Likelihood Ratiotest

$\underset{(r+1)\times 1}{\underset{\sim}{\beta}} = \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \hline \underset{\sim}{\beta}_2 \end{bmatrix} \begin{matrix} (q+1)\times 1 \\ \\ (r-q)\times 1 \end{matrix}$ $\underset{n\times(r+1)}{Z} = \underset{n\times(q+1)}{\left[Z_1 \right.} \underset{n\times(r-q)}{\left| Z_2 \right]}$.

Then $\underset{\sim}{y} = Z\underset{\sim}{\beta} + \underset{\sim}{\varepsilon} = \underset{n\times(q+1)}{\left[Z_1 \right.} \underset{n\times(r-q)}{\left| Z_2 \right]} \begin{bmatrix} \underset{\sim}{\beta}_1 \\ \hline \underset{\sim}{\beta}_2 \end{bmatrix} + \underset{\sim}{\varepsilon}$.

$= \underset{n\times 1}{Z_1}\underset{\sim}{\beta}_1 + \underset{n\times 1}{Z_2}\underset{\sim}{\beta}_2 + \underset{n\times 1}{\underset{\sim}{\varepsilon}}$

"Test"

$H_0 : \underset{\sim}{\beta}_2 = \underset{\sim}{0}$ vs $H_1 : \underset{\sim}{\beta}_2 \neq \underset{\sim}{0}$.

Recall The likelihood ratio test (LRT)

$\Lambda = \dfrac{\sup_{H_0} L(\underset{\sim}{\beta}, \sigma^2)}{\sup L(\underset{\sim}{\beta}, \sigma^2)}$ is the likelihood Ratio.

We reject H_0 if Λ is small.

We express Λ as a monotone function of a statistic T (of the data) whose null distribution (under H_0) is known.

Then $\{\Lambda < c\} \Leftrightarrow \{T < c' \text{ or } > c'\}$ for any constant c.

Then, for any constant c, there is another constant c' s.t.

$\{\Lambda < c\} \iff \{T < c'\}$ if Λ is increasing

$\iff \{T > c'\}$ if Λ is decreasing.

Then we will reject H_0 if $\begin{cases} T < c' & \text{in case 1} \\ T > c' & \text{in case 2.} \end{cases}$

Since the null distribution of T will be known, we can find c' so that the "level condition" is met.

In our case,

$$L(\beta, \sigma^2) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(Y - Z\beta)'(Y - Z\beta)\right]$$

The maximize of $L(\beta, \sigma^2)$ are

$$\hat{\beta} = (Z'Z)^{-1}Z'Y, \quad \hat{\sigma}^2 = \frac{(Y - Z\hat{\beta})'(Y - Z\hat{\beta})}{n}$$

$$\sup L(\beta, \sigma^2) = L(\hat{\beta}, \hat{\sigma}^2) = (2\pi)^{-n/2} (\hat{\sigma}^2)^{-n/2} e^{-n/2}$$

If H_0 ($\beta_2 = 0$) is true, then $Y = Z_1\beta_1 + \varepsilon$.

$$L\left(\begin{pmatrix} \beta_1 \\ 0 \end{pmatrix}, \sigma^2\right) = (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(Y - Z_1\beta_1)'(Y - Z_1\beta_1)\right]$$

The maximizer of $L(\beta_1, \sigma^2)$ are

$$\hat{\beta}_1 = (Z_1'Z_1)^{-1}Z_1'Y, \quad \hat{\sigma}_1^2 = \frac{(Y - Z_1\hat{\beta}_1)'(Y - Z_1\hat{\beta}_1)}{n}$$

So, $\sup_{H_0} L(\beta, \sigma^2) = L\left(\begin{pmatrix} \hat{\beta}_1 \\ 0 \end{pmatrix}, \hat{\sigma}_1^2\right)$

$$= (2\pi)^{-n/2} (\hat{\sigma}_1^2)^{-n/2} e^{-n/2}$$

So, $\Lambda = \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}^2}\right)^{-n/2} = \left(\frac{Y'(I - P_{Z_1})Y}{Y'(I - P_Z)Y}\right)^{-n/2}$

"decreasing"

So, we will reject H_0 if $\frac{Y'(I - P_{Z_1})Y}{Y'(I - P_Z)Y} > c$, for some c.

Recall that $(I - P_Z)$, $(I - P_{Z_1})$ are idempotent.

So, $\frac{Y'(I - P_Z)Y}{\sigma^2} \sim \chi^2_{n-r-1}$.

Under H_0, $\frac{Y'(I - P_{Z_1})Y}{\sigma^2} \sim \chi^2_{n-q-1}$. $\text{rank}(Z_1) = q+1$

"not independent"

These two χ^2 will be independent if $(I - P_Z)(I - P_{Z_1}) = 0$.

$$= I - P_Z - P_{Z_1} + P_{Z_1} \overset{?}{\neq} 0$$

"Not idependent"

But, ($\because P_Z \cdot Z_1 = Z_1$)

$$\frac{Y'(I - P_{Z_1})Y}{Y'(I - P_Z)Y} - 1 = \frac{Y'(P_Z - P_{Z_1})Y}{Y'(I - P_Z)Y}$$

independent.

Check $(P_Z - P_{Z_1})(I - P_Z) = P_Z - P_Z^2 - P_{Z_1} + P_{Z_1} = 0.$

Under H_0,

$$\frac{\frac{1}{r-q}\underset{\sim}{Y}'(P_Z - P_{Z_1})\underset{\sim}{Y}/\sigma^2}{\frac{1}{n-r-1}\underset{\sim}{Y}'(I-P_Z)\underset{\sim}{Y}/\sigma^2} \sim F_{r-q,\,n-r-1}$$

s^2

so, $\frac{1}{r-q}\frac{1}{s^2}\underset{\sim}{Y}'(P_Z - P_{Z_1})\underset{\sim}{Y} \sim F_{r-q,\,n-r-1}$.

So, we reject H_0 at level of α if

$$\underset{\sim}{Y}'(P_Z - P_{Z_1})\underset{\sim}{Y} > (r-q)s^2 F_{r-q,\,n-r-1}(\alpha)$$

» finv(1-α, r-q, n-r-1)

test Λ, decreasing, —

$\underset{\sim}{\beta} = \begin{bmatrix}\underset{\sim}{\beta}_1 \\ \underset{\sim}{\beta}_2\end{bmatrix}$. Want a 100(1-α)% CR for $\underset{\sim}{\beta}_2$.

* Confidence Region for $\underset{\sim}{\beta}_2$

Recall $\hat{\underset{\sim}{\beta}} \sim N_{r+1}(\underset{\sim}{\beta}, \sigma^2(Z'Z)^{-1})$. Then $\hat{\underset{\sim}{\beta}}_2 \sim$?

$$\hat{\underset{\sim}{\beta}}_2 \sim N_{r-q}\left(\underset{\sim}{\beta}_2, \sigma^2(Z_2'(I-P_{Z_1})Z_2)^{-1}\right)$$

$Z = [Z_1 \mid Z_2]$

$$Z'Z = \begin{bmatrix}Z_1' \\ Z_2'\end{bmatrix}[Z_1, Z_2] = \begin{bmatrix}Z_1'Z_1 & Z_1'Z_2 \\ Z_2'Z_1 & Z_2'Z_2\end{bmatrix}$$

Inverse of block matrix. $\begin{bmatrix}A & B \\ C & D\end{bmatrix}^{-1} = \begin{bmatrix}* & * \\ * & (D-CA^{-1}B)^{-1}\end{bmatrix}$

$$(Z'Z)^{-1} = \begin{bmatrix}* & * \\ * & (Z_2'Z_2 - Z_2'Z_1(Z_1'Z_1)^{-1}Z_1'Z_2)^{-1}\end{bmatrix} \quad (Z_2'(I-P_{Z_1})Z_2)^{-1}$$

$$\frac{(\hat{\underset{\sim}{\beta}}_2 - \underset{\sim}{\beta}_2)'Z_2'(I-P_{Z_1})Z_2(\hat{\underset{\sim}{\beta}}_2 - \underset{\sim}{\beta}_2)}{\sigma^2} \sim \chi^2_{r-q}$$

Then $\frac{1}{r-q}\frac{1}{s^2}(\hat{\underset{\sim}{\beta}}_2 - \underset{\sim}{\beta}_2)'Z_2'(I-P_{Z_1})Z_2(\hat{\underset{\sim}{\beta}}_2 - \underset{\sim}{\beta}_2) \sim F_{r-q,\,n-r-1}$

So, a 100(1-α)% CR for $\underset{\sim}{\beta}_2$ is

$$\left\{\underset{\sim}{\beta}_2 : (\hat{\underset{\sim}{\beta}}_2 - \underset{\sim}{\beta}_2)'Z_2'(I-P_{Z_1})Z_2(\hat{\underset{\sim}{\beta}}_2 - \underset{\sim}{\beta}_2) \le (r-q)s^2 F_{r-q,\,n-r-1}(\alpha)\right\}$$

H_0 is true ↑ ~~being~~ belong ~~to~~ in this ellipse.

• Prediction using the fitted regression model.

We have fitted a linear regression model on

$$\underset{\sim}{Y} = \begin{pmatrix}Y_1 \\ \vdots \\ Y_n\end{pmatrix}, \text{ and } Z = \begin{bmatrix}1 & Z_{11} & \cdots & Z_{1r} \\ \vdots & \vdots & & \vdots \\ 1 & Z_{n1} & \cdots & Z_{nr}\end{bmatrix}$$

For a future observations with known predictor variables.

(say $\underset{\sim}{Z}_0 = [1\ Z_{01} \cdots Z_{0r}]'$) how to obtain point and interval estimates of the corresponding (1) response and (2) mean response.

predictor

$$\underset{\sim}{Z}_0 = \begin{bmatrix}1 \\ Z_{01} \\ \vdots \\ Z_{0r}\end{bmatrix}$$

If Y_0 is the corresponding response, then

$Y_0 = \beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r} + \varepsilon_0$ $\quad \underset{\sim}{z}_0'\underset{\sim}{\beta} + \varepsilon_0 =$

$= \underset{\sim}{z}_0'\underset{\sim}{\beta} + \varepsilon_0$ $\quad (1\ z_{01}\ \cdots\ z_{0r}) \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_r \end{pmatrix} + \varepsilon_0$

$\mathbb{E}[Y_0] = \underset{\sim}{z}_0'\underset{\sim}{\beta}$. This is a parameter. We estimate this by

$\underset{\sim}{z}_0'\hat{\underset{\sim}{\beta}}$ (This is the BLUE). CI for $\underset{\sim}{z}_0'\underset{\sim}{\beta}$ is

$$\underset{\sim}{z}_0'\hat{\underset{\sim}{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2\underset{\sim}{z}_0'(\underset{\sim}{Z}'\underset{\sim}{Z})^{-1}\underset{\sim}{z}_0}$$

this is for an average ~~not~~ prediction.

To predict Y_0, the point estimate $\underset{\sim}{z}_0'\hat{\underset{\sim}{\beta}}$ is unbiased.

Interval estimate here is called prediction interval.

$\mathrm{var}(\hat{Y}_0) = \mathrm{var}(\underset{\sim}{z}_0'\hat{\underset{\sim}{\beta}}) + \sigma^2$

Prediction interval : $\underset{\sim}{z}_0'\hat{\underset{\sim}{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2 + s^2\underset{\sim}{z}_0'(\underset{\sim}{Z}'\underset{\sim}{Z})^{-1}\underset{\sim}{z}_0}$

Prediction Interval

Set up We have our regression model $\underset{\sim}{Y} = \underset{\sim}{Z}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$ (number; parameters; response; predictors; error)

We observe values of y and $\underset{\sim}{Z}$ for n observations.

We arrange them to have

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \quad \underset{\sim}{Z} = \begin{pmatrix} \underset{\sim}{z}_1' \\ \vdots \\ \underset{\sim}{z}_n' \end{pmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1r} \\ \vdots & \vdots & & \\ 1 & z_{n1} & & z_{nr} \end{bmatrix}$$

Using these we estimate $\underset{\sim}{\beta}$: $\hat{\underset{\sim}{\beta}} = (\underset{\sim}{Z}'\underset{\sim}{Z})^{-1}\underset{\sim}{Z}'\underset{\sim}{Y}$

Now, suppose we have a future observation for which the predictor variables are known.

But, we do not know the value of the response.

$Y_0 = \underset{\sim}{z}_0'\underset{\sim}{\beta} + \varepsilon_0$ (same as earlier; error for future obs.)

Y_0: response for future obs. $\quad \underset{\sim}{z}_0$: predictor for future obs

$\mathbb{E}(Y_0) = \underset{\sim}{z}_0'\underset{\sim}{\beta}$ (mean responses.) This is a parameter, so we can find estimate and CI for it.

• Prediction Interval

Goal: to find an interval that will contain Y_0 (response of a future observation) with a given confidence.

Consider $Y_0 - z_0'\hat{\beta}$ and find its distribution.

Recall, under normality assumption, $\hat{\beta} \sim N_{r+1}(\beta, \sigma^2(Z'Z)^{-1})$.

Then $z_0'\hat{\beta} \sim N(z_0'\beta, \sigma^2 z_0'(Z'Z)^{-1}z_0)$

($z_0'\hat{\beta}$: scalar; σ^2: variance from ε_0 (future error))

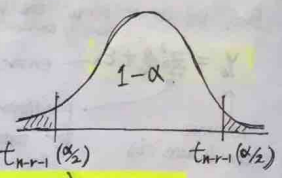
ind.

$Y_0 \sim N(z_0'\beta, \sigma^2)$

So, $Y_0 - z_0'\hat{\beta} \sim N(0, \sigma^2(1 + z_0'(Z'Z)^{-1}z_0))$

$$\frac{Y_0 - z_0'\hat{\beta}}{\sqrt{\sigma^2(1+z_0'(Z'Z)^{-1}z_0)}} \sim N(0,1)$$

$$\frac{Y_0 - z_0'\hat{\beta}}{\sqrt{s^2(1+z_0'(Z'Z)^{-1}z_0)}} \sim t_{n-r-1}$$



So, $P\left(\left|\frac{Y_0 - z_0'\hat{\beta}}{\sqrt{s^2(1+z_0'(Z'Z)^{-1}z_0)}}\right| \le t_{n-r-1}\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$

So, 100 $(1-\alpha)$% PI for Y_0

$$z_0'\hat{\beta} \pm \sqrt{s^2(1+z_0'(Z'Z)^{-1}z_0)}\, t_{n-r-1}\left(\frac{\alpha}{2}\right)$$

• Model verification

Recall, $\hat{\varepsilon}_j = Y_j - \hat{y}_j$

$$\hat{\varepsilon} = \begin{bmatrix}\hat{\varepsilon}_1 \\ \vdots \\ \hat{\varepsilon}_n\end{bmatrix} = Y - \hat{Y} = (I - P_Z)Y$$

$$\mathrm{Var}(\hat{\varepsilon}) = \sigma^2(I - P_Z)$$

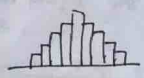
Normalized error estimate

$$\hat{\varepsilon}_j^x = \frac{\varepsilon_j}{\sqrt{\widehat{\mathrm{var}}(\hat{\varepsilon}_j)}} = \frac{\varepsilon_j}{s^2(I-P_Z)_{j,j}}$$

~~normalized variance~~.

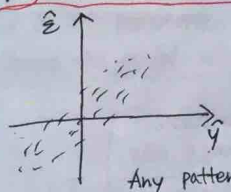
• Residual Plot

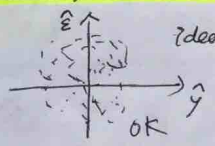
If we plot the histogram of $\hat{\varepsilon}_1^x, \ldots, \hat{\varepsilon}_n^x$, then we expect the histogram to be close to that of $N(0,1)$



• Violations indicate violations of normality assumption on

Plot $(\hat{y}_1, \hat{\varepsilon}_1), \ldots, (\hat{y}_n, \hat{\varepsilon}_n)$



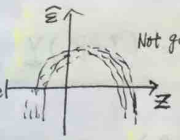


Any pattern in this plot indicates that the Lin. model assumption is not true.

Plot $\hat{\underset{\sim}{\varepsilon}}$ against $Z_{*2}, \cdots, Z_{*(r+1)}$

Expect no pattern if assumption are right.

But there is pattern, that indicates violation of the linear model namely corresponding Z variable needs to be transformed before using.

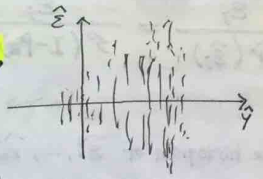


Not good.

~~Volatility plot~~

In the plot $\hat{\underset{\sim}{\varepsilon}}$ vs $\hat{\underset{\sim}{Y}}$,

such a pattern violates the "equal variance" assumption.



~~Q-Q plot~~

Q-Q plot for the normalized residuals

To check "normality assumption".

Model selections (predictors)

When the number of predictors is large, we need to select "suitable small" subset of predictors.

First, naive approach is to consider all subsets and choose the one which maximizes R^2 (rss)

This doesn't work, as R^2 is an increasing function of r (# of predictors).

Adjusted R^2, $\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-r-1}$

One method is to use $\bar{R}^2$ to select model.

A "better" measure is Mallow's

$$C_p = \frac{\text{residual SS using predictors}}{\text{residual variances } (\sigma^2)} - (n-2p)$$

p = # variables including 1

n = # observations.

C_p is not monotone.

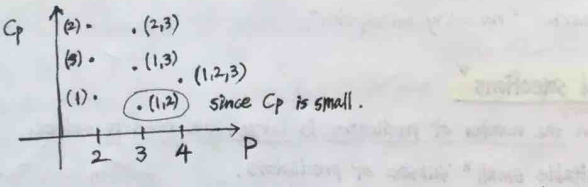
Choose the model with minimum C_p.

$$C_p = \left(\frac{\text{residual sum of squares for subset model with p parameters, including an intercept}}{\text{residual variance for full model (residual SS)}}\right) - (n-2p)$$

$p = r+1$

A plot of the pairs (p, C_p), one for each subset of predictors.

Ex suppose $Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \beta_3 Z_3 + \varepsilon$



These two methods work when # predictors is moderate, as we need to consider all possible subsets.

Next, we will see step-wise regression.

Math B7800 2/27/18, Tue

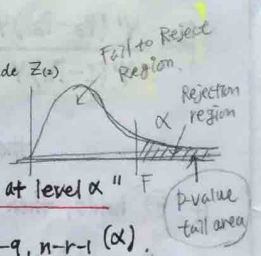
- Predictor selection
- Stepwise selection method

Recall

If $\beta = \begin{pmatrix} \beta_{(1)} \\ \beta_{(2)} \end{pmatrix} \begin{matrix} (q+1)\times 1 \\ (r-q)\times 1 \end{matrix}$

Test $H_0 : \beta_{(2)} = 0$ vs $H_1 : \beta_{(2)} \neq 0$ "likelihood ration test"

$\underset{n\times(r+1)}{Z} = \left[\underset{n\times(q+1)}{Z_{(1)}} \mid \underset{n\times(r-q)}{Z_{(2)}} \right]$ if H_0, exclude $Z_{(2)}$. if H_1, we need to include $Z_{(2)}$



we use the LRT which boils down to

$$F = \frac{\underset{\sim}{Y}'(P_Z - P_{Z_1})\underset{\sim}{Y}\frac{1}{r-q}}{\underset{\sim}{Y}'(I - P_Z)\underset{\sim}{Y}\frac{1}{n-r-1}}$$

"we reject H_0 at level α" when $F > F_{r-q,\, n-r-1}(\alpha)$.

We will refer to the model: $Y = \beta_0 + \beta_1 Z_1 + \cdots + \beta_r Z_r + \varepsilon$ as "regression function"

- Stepwise method

Step 1: Scan through all the variables (one at time) and which contributes the most to the Regression SS.

$Y = \beta_0 + \beta_1 Z_1$
$Y = \beta_0 + \beta_2 Z_2$
$\vdots$
$Y = \beta_0 + \beta_r Z_r$

→ Find which one has R^2 value the most.

$Z = \begin{bmatrix} 1 & Z_{11} \\ \vdots & \vdots \\ 1 & Z_{n1} \end{bmatrix}$ → compute $\frac{Y'(P_Z - P_1)Y}{Y'(I - P_1)Y}$ * use AIC, $\bar{R}^2$, Cp

$Z = \begin{bmatrix} 1 & Z_{1r} \\ \vdots & \vdots \\ 1 & Z_{nr} \end{bmatrix}$

Suppose k is the one which maximizes

Step 2 Test $\beta = \begin{bmatrix} \beta_0 \\ \beta_k \end{bmatrix}$

Test $H_0: \beta_k = 0$ vs $H_1: \beta_k \neq 0$

$$Z = \begin{bmatrix} 1 & z_{1k} \\ \vdots & \vdots \\ 1 & z_{nk} \end{bmatrix} \quad Z_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$F = \frac{Y'(P_Z - P_{Z_1})Y \ \frac{1}{1-0}}{Y'(I-P_Z)Y \ \frac{1}{n-1-1}} > F_{1,n-2}(\alpha) \quad -(*) \qquad \{r=1, q=0$$

$\Downarrow$ reject $H_0 \Rightarrow \beta_k \neq 0$

then include k in the model.

if $(*)$ holds, then "include" Z_k in the regression function.

Current model is $Y = \beta_0 + \beta_k Z_k + \varepsilon$.

Step 3 We scan through the remaining variables (one at time) and find out which one contributes to the Reg. SS the most. (keeping the existing regression function)

Current model $Y = \beta_0 + \beta_k Z_k + \beta_1 Z_1 \rightarrow Z = \begin{bmatrix} 1 & z_{1k} & z_{11} \\ \vdots & \vdots & \vdots \\ 1 & z_{nk} & z_{n1} \end{bmatrix}, \ Z_1 = \begin{bmatrix} 1 & z_{1k} \\ \vdots & \vdots \\ 1 & z_{nk} \end{bmatrix}$

$\vdots$

$Y = \beta_0 + \beta_k Z_k + \beta_{k-1} Z_{k-1}$

$Y = \beta_0 + \beta_k Z_k + \beta_{k+1} Z_{k+1} \rightarrow Z = \begin{bmatrix} 1 & z_{1k} & z_{1r} \\ \vdots & \vdots & \vdots \\ 1 & z_{nk} & z_{nr} \end{bmatrix}$

$\vdots$

$Y = \beta_0 + \beta_k Z_k + \beta_r Z_r$.

Find out which model has "max value" of $Y'(P_Z - P_1)Y$. reg. SS

Suppose L is the one that maximizes:

Then $\beta = \begin{bmatrix} \beta_0 \\ \beta_k \\ \beta_L \end{bmatrix}$ $H_0: \beta_L = 0$ vs $H_1: \beta_L \neq 0$.

$$F = \frac{Y'(P_Z - P_{Z_1})Y \ \frac{1}{2-1}}{Y'(I-P_Z)Y \ \frac{1}{n-2-1}} \qquad \begin{matrix} r=2 \\ k=1 \\ q=1 \end{matrix}$$

$$Z = \begin{bmatrix} 1 & z_{1k} & z_{1L} \\ \vdots & \vdots & \vdots \\ 1 & z_{nk} & z_{nL} \end{bmatrix}, \ Z_1 = \begin{bmatrix} 1 & z_{1k} \\ \vdots & \vdots \\ 1 & z_{nk} \end{bmatrix}$$

Include Z_L if $F > F_{1,n-3}(\alpha)$

If included, then current model is

$Y = \beta_0 + \beta_k Z_k + \beta_L Z_L + \varepsilon$

Step 4 When a new variable is included, we need to check whether one of the existing ones fail the F test or not.

Suppose at a stage, we include Z_{im} to $(Z_{i1}, Z_{i2}, \ldots, Z_{im-1})$

then test:

$\beta_0 = \begin{bmatrix} \beta_0 \\ \beta_{i1} \\ \vdots \\ \beta_{im} \end{bmatrix}$

$H_0: \beta_{i1} = 0$ vs $H_1: \beta_{i1} \neq 0$ ($\beta_k = 0$, $\beta_k \neq 0$; Z_k, Z_L, ...)

$\vdots$

$H_0: \beta_{im-1} = 0$ vs $H_1: \beta_{im-1} \neq 0$.

"Repeat Steps 3,4"

If in one of the tests, the H_0 ($\beta_K = 0$) is accepted, then the corresponding variable goes out.

At the end, we will have a subset $\{i_1, \dots, i_p\}$ of $\{1, \dots, r\}$ such that no additional variable can be included, (i.e., F test is insignificant) and no existing variable leaves (i.e., the corresponding F test is significant). Then we stop.

accept H_0 → exclude the corresponding variable.

"don't drop." ⇔ reject H_0 ⇒ include the corresponding variable.

$$AIC = n \ln\left(\frac{\text{residual SS for subset model with } p \text{ parameters, including an intercept}}{n}\right) + 2p$$

overall, we want to select models from those having the smaller values of AIC.

• Information Criterion Method

The famous information criterion assigned to subsets of variables is the Akaike's Information on Criterion (AIC).

For a subset of size p,

$$AIC = n \log\left(\frac{\text{Residual SS for that subset}}{n}\right) + 2p$$

($r+1$ multiplied)

decreasing; increasing

one needs to minimize AIC to select variables.

• Multi-colinearity

So, far we have assumed that Z has full rank.

Then $Z'Z$ was nonsingular.

If this condition does not hold, then we say that the data has Multi-colinearity.

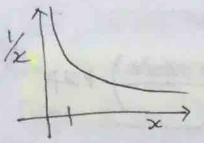
In some cases, $Z\underset{\sim}{a} = \underset{\sim}{c}$ for some $\underset{\sim}{a} \neq \underset{\sim}{0}$.

In this case, $Z'Z$ is singular.

In many cases, $Z'Z$ is "nearly singular", meaning it has some "very small" eigenvalues.

In the first case, we choose a linearly independent subset of columns of Z.

In the second case, $(Z'Z)^{-1}$ becomes numerically unstable.

e.g. 

0.01 → 100

0.02 → 50.

Then remove one of the highly correlated columns of Z.

exam 1.

3/1/18, Thur

⑨ Multivariate "Multiple" Regression.

For each individual, we have a vector of response $(1\times m)$

The model

$$\underset{\sim}{Y}' = [Y_1, \dots, Y_m] = [1, z_1, \dots, z_r] \begin{bmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \vdots & \ddots & \\ \beta_{r1} & \cdots & \beta_{rm} \end{bmatrix} + \underset{\sim}{\varepsilon}'$$

"m-responses" (multiple responses); $1\times(r+1)$; $(r+1)\times m$; $1\times m$

$$\mathbb{E}[\underset{\sim}{\varepsilon}] = \underset{m\times 1}{\underset{\sim}{0}}, \quad \mathbb{E}[\underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'] = \Sigma_{m\times m}$$

We observe the responses and predictors from n individuals.

values of

$$\underset{\sim}{Y}_1' = \underset{\sim}{z}_1'\beta + \underset{\sim}{\varepsilon}_1'$$
$$\vdots$$
$$\underset{\sim}{Y}_n' = \underset{\sim}{z}_n'\beta + \underset{\sim}{\varepsilon}_n'$$

We organize them:

$$Y = \begin{bmatrix} \underset{\sim}{Y}_1' \\ \vdots \\ \underset{\sim}{Y}_n' \end{bmatrix}_{n\times m}, \quad Z = \begin{bmatrix} \underset{\sim}{z}_1' \\ \vdots \\ \underset{\sim}{z}_n' \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \cdots & z_{nr} \end{bmatrix}_{n\times(r+1)}$$

$$\beta = \begin{bmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \vdots & \ddots & \vdots \\ \beta_{r1} & \cdots & \beta_{rm} \end{bmatrix}_{(r+1)\times m}, \quad \underset{n\times m}{\varepsilon} = \begin{bmatrix} \underset{\sim}{\varepsilon}_1' \\ \vdots \\ \underset{\sim}{\varepsilon}_n' \end{bmatrix} = \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1m} \\ \vdots & \ddots & \\ \varepsilon_{n1} & \cdots & \varepsilon_{nm} \end{bmatrix}$$

Multivariate multiple regression model: $Y = Z\beta + \varepsilon$

Each observation has m-response variables which are correlated.

$$\underset{1\times m}{\underset{\sim}{Y}'} = \underset{1\times(r+1)}{\underset{\sim}{Z}'}\underset{(r+1)\times m}{\beta} + \underset{1\times m}{\underset{\sim}{\varepsilon}'} \quad \text{(the model)}$$

$\mathbb{E}[\underset{\sim}{\varepsilon}'] = \underset{\sim}{0}'$

$\text{cov}(\underset{\sim}{\varepsilon}') = \Sigma_{m\times m} = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & & \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{bmatrix}$

"m = # responses for each individual observation"

We observe the values of the predictors and responses from n individuals.

We arrange them into $\underset{n\times m}{Y} = \begin{bmatrix} \underset{\sim}{Y}_1' \\ \vdots \\ \underset{\sim}{Y}_n' \end{bmatrix} \begin{matrix} [Y_{11},\ldots,Y_{1m}] \\ \vdots \\ [Y_{n1},\ldots,Y_{nm}] \end{matrix}$

$$\underset{n\times(r+1)}{Z} = \begin{bmatrix} \underset{\sim}{Z}_1' \\ \vdots \\ \underset{\sim}{Z}_n' \end{bmatrix} = \begin{bmatrix} 1 & Z_{11} & \cdots & Z_{1r} \\ \vdots & \vdots & & \vdots \\ 1 & Z_{n1} & & Z_{nr} \end{bmatrix}, \quad \underset{(r+1)\times m}{\beta} = \begin{bmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \vdots & & \vdots \\ \beta_{r1} & & \beta_{rm} \end{bmatrix}$$

$$\underset{n\times m}{\varepsilon} = \begin{bmatrix} \underset{\sim}{\varepsilon}_1' \\ \vdots \\ \underset{\sim}{\varepsilon}_n' \end{bmatrix} \begin{matrix} [\varepsilon_{11},\ldots,\varepsilon_{1m}] \\ \vdots \\ [\varepsilon_{n1},\ldots,\varepsilon_{nm}] \end{matrix}$$

So, the model becomes $Y = Z\beta + \varepsilon$

$\underset{\sim}{\varepsilon}_1', \ldots, \underset{\sim}{\varepsilon}_n'$ are independent.

$$Y = \left[\underset{\sim}{Y}_{(1)} \mid \cdots \mid \underset{\sim}{Y}_{(m)}\right], \quad \beta = \left[\underset{\sim}{\beta}_{(1)} \mid \cdots \mid \underset{\sim}{\beta}_{(m)}\right]$$

$$\varepsilon = \left[\underset{\sim}{\varepsilon}_{(1)} \mid \cdots \mid \underset{\sim}{\varepsilon}_{(m)}\right]$$

So, $\left[\underset{\sim}{Y}_{(1)} \mid \cdots \mid \underset{\sim}{Y}_{(m)}\right] = Z\left[\underset{\sim}{\beta}_{(1)} \mid \cdots \mid \underset{\sim}{\beta}_{(m)}\right] + \left[\underset{\sim}{\varepsilon}_{(1)} \mid \cdots \mid \underset{\sim}{\varepsilon}_{(m)}\right]$.

So, $\underset{\sim}{Y}_{(i)} = Z\underset{\sim}{\beta}_{(i)} + \underset{\sim}{\varepsilon}_{(i)}$, for $i = 1, 2, \ldots, m$.

Goal: Estimate β, i.e., $\hat{\beta}$.

One method is to estimate $\underset{\sim}{\beta}_{(i)}$ by $\hat{\underset{\sim}{\beta}}_{(i)}$ (the least square estimate from the univariate case).

$$\hat{\underset{\sim}{\beta}}_{(i)} = (Z'Z)^{-1}Z'\underset{\sim}{Y}_{(i)}$$

Using this, $\hat{\beta} = \left[\hat{\underset{\sim}{\beta}}_{(1)} \mid \cdots \mid \hat{\underset{\sim}{\beta}}_{(m)}\right]$

$$= (Z'Z)^{-1}Z'\left[\underset{\sim}{Y}_{(1)} \mid \cdots \mid \underset{\sim}{Y}_{(m)}\right]$$

$$= \underset{n\times m}{(Z'Z)^{-1}Z'Y}$$

For this $\hat{\beta}$, $\hat{Y} = Z\hat{\beta} = \underbrace{Z(Z'Z)^{-1}Z'}_{P_Z}Y = P_Z Y$.

and $\hat{\varepsilon} = Y - \hat{Y} = (I - P_Z)Y$

$\hat{Y} = P_Z Y$, $P_Z = Z(Z'Z)^{-1}Z'$

least square The error for an estimator B of $\hat{\beta}$ defined to (least square)

$$f(B) = tr\left[\underbrace{(Y - \hat{Y})'}_{m\times n}\underbrace{(Y - \hat{Y})}_{n\times m}\right]_{m\times m}$$

$\hat{Y} = ZB$

$A = [A_{(1)} \cdots A_{(m)}]$

$(A'A)_{i,j}$

Fact $\hat{\beta}$ minimizes $f(\ )$, i.e.,

$$f(\hat{\beta}) \le f(B) \quad \text{for any } B.$$

$$f(B) = \sum_{i=1}^{m} (\underset{\sim}{Y}_{(i)} - Z\underset{\sim}{\beta}_{(i)})'(\underset{\sim}{Y}_{(i)} - Z\underset{\sim}{\beta}_{(i)})$$

The i^{th} summand is minimized at $\underset{\sim}{B}_{(i)} = \hat{\underset{\sim}{\beta}}_{(i)}$ (from our knowledge about the univariate case).

So, $f(B) \ge \sum_{i=1}^{m} (\underset{\sim}{Y}_{(i)} - Z\hat{\underset{\sim}{\beta}}_{(i)})'(\underset{\sim}{Y}_{(i)} - Z\hat{\underset{\sim}{\beta}}_{(i)}) = f(\hat{\beta})$

Now, suppose we consider a different least square error

"$g(B) = |(Y-ZB)'(Y-ZB)|$"

Fact (from Lin. Alg.)

If (C) and (D) are symmetric non-negative definite (nnd) matrices of same size, then $|C+D| \ge |C| + |D|$

determinate = product of eigenvalue

determinant, not absolute value.

proof case 1: If C and D are both positive semi definite (nnd but not pd) (some eigenvalue is zero), then $|C| = |D| = 0$ and $|C+D| \ge 0$

case 2: one of the matrices is pd. say D is pd.

Then $C+D = D^{1/2}(D^{-1/2}CD^{-1/2})D^{1/2} + D$.

$= D^{1/2}(D^{-1/2}CD^{-1/2} + I)D^{1/2}$

For any nnd matrix A, $|I+A| \ge 1+|A|$

So, $|C+D| = |D^{1/2}|\underbrace{|I + D^{-1/2}CD^{-1/2}|}_{\ge 1+|D^{-1/2}CD^{-1/2}|}\cdot|D|^{1/2} \ge |D^{1/2}|(1+|D^{-1/2}CD^{-1/2}|)|D^{1/2}|$

$= |D| + |C|$ $|D^{1/2}| = C$

write $Y - ZB = (Y - Z\hat{\beta}) + Z(\hat{\beta} - B)$.

So, $(Y-ZB)'(Y-ZB)$

$= (Y-Z\hat{\beta})'(Y-Z\hat{\beta}) + (Z(\hat{\beta}-B))'Z(\hat{\beta}-B)$

$+ \underbrace{(Y-Z\hat{\beta})'Z}_{=0}(\hat{\beta}-B) + [Z(\hat{\beta}-B)]'(Y-Z\hat{\beta})$

↖ by $\hat{\beta} = (Z'Z)^{-1}Z'Y \Leftrightarrow (Z'Z)\hat{\beta} = Z'Y$

$(\hat{\beta}-B)'\underbrace{Z'(Y-Z\hat{\beta})}_{=0}$

So, $g(B) = |(Y-ZB)'(Y-ZB)|$

$= |(Y-Z\hat{\beta})'(Y-Z\hat{\beta}) + (Z\hat{\beta}-ZB)'(Z\hat{\beta}-ZB)|$

$\ge |(Y-Z\hat{\beta})'(Y-Z\hat{\beta})| + |(Z\hat{\beta}-ZB)'(Z\hat{\beta}-ZB)|$

So, $g(B) \ge g(\hat{\beta}) + |(\hat{\beta}-B)'Z'Z(\hat{\beta}-B)|$

$\ge g(\hat{\beta})$ (as A'A is nnd for any A).

$\Rightarrow |A'A|$ is non-negative.

Under the additional normality assumption, i.e., $\underset{\sim}{\varepsilon}' \sim N_m(\underset{\sim}{0}, \Sigma)$ (unknown) we want to find MLE for β, Σ.

Def of MLE, likelihood func...

The likelihood function

$L(\beta, \Sigma)$ = joint density of Y's

$= \prod_{i=1}^{n} (\text{pdf of } Y_i')$

Recall $Y' = Z'\beta + \varepsilon'$

If $\varepsilon' \sim N_m(0, \Sigma)$, then $Y' \sim N_m(Z'\beta, \Sigma)$.

pdf of Y: $f_Y(y) = (2\pi)^{-m/2} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}\underbrace{(y'-z'\beta)}_{(y-\beta'z)'}\Sigma^{-1}(y-\beta'z)\right)$

$= \prod_{i=1}^{n} f_{Y_i}(y_i) = (2\pi)^{-\frac{nm}{2}} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta'z_i)'\Sigma^{-1}(y_i-\beta'z_i)\right)$

So, $L(\beta,\Sigma) = (2\pi)^{-\frac{nm}{2}} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\underset{n\times m}{\underset{|}{tr}}\left[\underset{n\times m}{(Y-Z\beta)}\underset{m\times m}{\Sigma^{-1}}\underset{m\times n}{(Y-Z\beta)'}\right]\right)$

Fact by $tr(CD) = tr(DC)$

$= (2\pi)^{-\frac{nm}{2}} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} tr\left[\Sigma^{-1}(Y-Z\beta)'(Y-Z\beta)\right]\right)$

We will maximize in 2 steps.

First, wrt Σ, then wrt β.

Step 1 β is fixed. Call $(Y-Z\beta)'(Y-Z\beta) = D$.

define $h(\Sigma) = |\Sigma|^{-n/2} \exp\left(-\frac{1}{2} tr(\Sigma^{-1}D)\right)$

$tr(\Sigma^{-1}D) = tr(\Sigma^{-1/2}D\Sigma^{-1/2}) = \sum_{i=1}^{m}\lambda_i$, where $\lambda_1, \ldots, \lambda_m$ are the eigenvalues of $\Sigma^{-1/2}D\Sigma^{-1/2}$.

Then, $\prod_{i=1}^{m}\lambda_i = \left|\Sigma^{-1/2}D\Sigma^{-1/2}\right| = \frac{|D|}{|\Sigma|}$

then, $h(\Sigma) = \frac{1}{|D|^{n/2}}\left(\frac{|D|}{|\Sigma|}\right)^{n/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\lambda_i\right)$

$= \frac{1}{|D|^{n/2}} \prod_{i=1}^{m}\lambda_i \, e^{-\frac{1}{2}\lambda_i}$

A is nnd if $\underset{\sim}{x}'A\underset{\sim}{x} \geq 0$ for all $\underset{\sim}{x}$

If A is symmetric, then nnd $\Leftrightarrow$ all eigen values ≥ 0

A is p.d if $\underset{\sim}{x}'A\underset{\sim}{x} > 0$ unless $\underset{\sim}{x} = \underset{\sim}{0}$.

if A is symmetric, then pd $\Leftrightarrow$ all eig values > 0.

positive semidefinite = {nnd} − {pd}.

Recall multivariate multiple linear regression.

model $\underset{n\times m}{Y} = \underset{n\times(r+1)}{Z}\ \underset{(r+1)\times m}{\beta} + \underset{n\times m}{\varepsilon}$

$\hat{\beta} = (Z'Z)^{-1}Z'Y$

$\hat{\Sigma} = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon}$

Result

If the rows of ε are iid $N_m(\underset{\sim}{0}, \Sigma)$, then the MLE of β and Σ will be $\hat{\beta} = (Z'Z)^{-1}Z'Y$, $\hat{\Sigma} = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon}$, where

$\hat{\underset{\sim}{\varepsilon}} = Y - \hat{Y} = Y - Z\hat{\beta} = Y - Z(Z'Z)^{-1}Z'Y = (I - P_Z)Y$.

Proof Recall the likelihood function

$L(\beta, \Sigma) = (2\pi)^{\frac{-nm}{2}} |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}D)\right)$, where

$D = (Y - Z\beta)'(Y - Z\beta)$

For fixed β (so, D is also fixed), we will maximize

$f(\Sigma) = |\Sigma|^{-n/2} \exp\left(-\frac{1}{2}\operatorname{tr}(\Sigma^{-1}D)\right)$.

$= \frac{1}{|D|^{n/2}}\left(\frac{|D|}{|\Sigma|}\right)^{n/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\Sigma^{-1/2} D \Sigma^{-1/2})\right]$

symmetric, then eigenvalues are real value.

If the eigenvalues of $\Sigma^{-1/2} D \Sigma^{-1/2}$ are $\lambda_1, \dots, \lambda_n$,

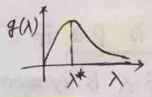
then $tr(\Sigma^{-1/2} D \Sigma^{-1/2}) = \sum_{i=1}^{m} \lambda_i$

$$\prod_{i=1}^{m} \lambda_i = \left| \Sigma^{-1/2} D \Sigma^{-1/2} \right| = |\Sigma^{-1/2}|\,|D|\,|\Sigma^{-1/2}| = \frac{|D|}{|\Sigma|}$$

("determinant")

Then

$$f(\Sigma) = \frac{1}{|D|^{n/2}} \left(\prod_{i=1}^{m} \lambda_i \right)^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{m} \lambda_i \right)$$



$$= \frac{1}{|D|^{n/2}} \prod_{i=1}^{m} g(\lambda_i), \text{ where } g(\lambda) = \lambda^{n/2} e^{-\lambda/2}$$

(zero & positive eigenvalues)

We will maximize g wrt. λ. If λ^* maximizes g, then $f(\Sigma)$ will be maximized by taking $\lambda_i = \lambda^*$ for all $i = 1, 2, \dots, m$.

$$\frac{d}{d\lambda} \log(g(\lambda)) = \frac{n}{2\lambda} - \frac{1}{2}$$

(concave function; $\frac{n}{2}\log\lambda - \frac{\lambda}{2}$)

so, $\frac{d}{d\lambda}(\log g(\lambda)) = 0 \iff \lambda = n$, so $\lambda^* = n$.

To find maximizer of $f(\Sigma)$, we need to find Σ so that all eigenvalues of $\Sigma^{-1/2} D \Sigma^{-1/2}$ are equal to n.

Thus, we must have $\Sigma^{-1/2} D \Sigma^{-1/2} = nI \Rightarrow n\Sigma = D$

("want"; multiply $\Sigma^{-1/2}$ to both sides.)

$$\Sigma = \frac{1}{n} D$$

so, $f(\Sigma) \le f(\frac{1}{n} D)$

so, $L(\beta, \Sigma) \le L(\beta, \frac{1}{n} D) = L(\beta, \frac{1}{n}(Y - Z\beta)'(Y - Z\beta))$

$$= (2\pi)^{\frac{-nm}{2}} \left| \tfrac{1}{n} D \right|^{-n/2} \exp\left[-\tfrac{1}{2} tr\left((\tfrac{1}{n} D)^{-1} D \right) \right]$$

$$= (2\pi)^{\frac{-nm}{2}} |D|^{-n/2} n^{n/2} e^{-\frac{mn}{2}}$$

$$\le L\left(\hat{\beta}, \tfrac{1}{n}(Y - Z\hat{\beta})'(Y - Z\hat{\beta}) \right)$$

because we have already seen that $\beta = \hat{\beta}$ minimizes $h(\beta) = |(Y - Z\beta)'(Y - Z\beta)|$

$$\max_{\beta, \Sigma} L(\beta, \Sigma) = c \left| \hat{\Sigma} \right|^{-n/2}, \quad c \text{ is constant}$$

$$(2\pi)^{-nm/2} n^{n/2} e^{\frac{-nm}{2}}$$

properties of $\hat{\beta}$.

$$E[\hat{\beta}] = E[(Z'Z)^{-1}Z'Y] = (Z'Z)^{-1}Z' \underbrace{E(Y)}_{Z\beta} = (Z'Z)^{-1}Z'Z\beta = \beta.$$

Recall $Y = [Y_{(1)} | \cdots | Y_{(m)}] = Z\beta + \varepsilon$

$E[\hat{\beta}_{(i)}] = \beta_{(i)}$, for $i = 1, \dots, m$

$= Z[\beta_{(1)} | \cdots | \beta_{(m)}] + [\varepsilon_{(1)} | \cdots | \varepsilon_{(m)}]$

fact: $(CD)_{xj} = C D_{xj}$.

$\hat{\beta} = (Z'Z)^{-1}Z'Y$

so, $\hat{\beta}_{(j)} = (Z'Z)^{-1}Z'Y_{(j)}$.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & & \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{bmatrix}$$

$m \times m$ (# of responses)

$\star$ $Cov(\hat{\beta}_{(j)}, \hat{\beta}_{(k)}) =$

$Cov((Z'Z)^{-1}Z'Y_{(j)}, (Z'Z)^{-1}Z'Y_{(k)})$

$= (Z'Z)^{-1}Z' \underbrace{Cov(Y_{(j)}, Y_{(k)})}_{= Cov(\varepsilon_{(j)}, \varepsilon_{(k)}) = \sigma_{jk} I_{n \times n}} Z (Z'Z)^{-1}$

$$\varepsilon_{(j)} = \begin{bmatrix} \varepsilon_{j1} \\ \vdots \\ \varepsilon_{jn} \end{bmatrix}, \quad \varepsilon_{(k)} = \begin{bmatrix} \varepsilon_{k1} \\ \vdots \\ \varepsilon_{kn} \end{bmatrix}$$

$= (Z'Z)^{-1}Z'(\sigma_{jk} I_n) Z (Z'Z)^{-1}$

$= \sigma_{jk} (Z'Z)^{-1}$

Also, $\hat{Y}'\hat{\varepsilon} = (P_Z Y)'(I-P_Z)Y = Y' \underbrace{P_Z(I-P_Z)}_{=0} Y$

$\text{cov}(\hat{y}_{(j)}, \hat{\varepsilon}_{(k)}) = 0$ for j, k.

$= \text{cov}(P_Z Y_{(j)}, (I-P_Z)Y_{(k)})$

$= P_Z \,\text{cov}(Y_{(j)}, Y_{(k)})\,(I-P_Z) = P_Z(\sigma_{jk} I)(I-P_Z) = 0.$

Recall $\hat{\Sigma} = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n}(Y-Z\hat{\beta})'(Y-Z\hat{\beta})$

$\mathbb{E}[\hat{\Sigma}] =$ (to check ↑ unbiased)

$\hat{\varepsilon} = [\hat{\varepsilon}_{(1)} | \cdots | \hat{\varepsilon}_{(m)}]$

$\hat{\varepsilon}'\hat{\varepsilon} = \begin{bmatrix} \hat{\varepsilon}'_{(1)}\hat{\varepsilon}_{(1)} & \cdots & \hat{\varepsilon}'_{(1)}\hat{\varepsilon}_{(m)} \\ \vdots & & \\ \hat{\varepsilon}'_{(m)}\hat{\varepsilon}_{(1)} & \cdots & \hat{\varepsilon}'_{(m)}\hat{\varepsilon}_{(m)} \end{bmatrix}$

Find $\mathbb{E}[\hat{\varepsilon}'_{(j)}\hat{\varepsilon}_{(k)}]$

$= \mathbb{E}[Y'_{(j)}(I-P_Z)^2 Y_{(k)}]$

$= \mathbb{E}[Y'_{(j)}(I-P_Z)Y_{(k)}] = \mathbb{E}[\text{tr}(Y'_{(j)}(I-P_Z)Y_{(k)})]$

$= \mathbb{E}[\text{tr}((I-P_Z)Y_{(k)}Y'_{(j)})] = \text{tr}\,\mathbb{E}[(I-P_Z)Y_{(k)}Y'_{(j)}]$

$= \text{tr}\left((I-P_Z)\left\{\text{cov}(Y_{(k)}, Y'_{(j)}) + Z\beta_{(k)}\beta'_{(j)}Z'\right\}\right)$

$= \text{tr}[(I-P_Z)\sigma_{jk} I + 0]$ (" $\text{tr}(I-P_Z)$.)

$= \sigma_{jk}\,\text{tr}(I-P_Z) = \sigma_{jk}\,\text{rank}(I-P_Z)$ (↑ idempotent matrix)

$= \sigma_{jk}(n-r-1)$

So, $\mathbb{E}[\hat{\varepsilon}'\hat{\varepsilon}] = (n-r-1)\Sigma$

So, unbiased estimator of Σ is

$S = \frac{1}{n-r-1}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{n-r-1}(Y-Z\hat{\beta})'(Y-Z\hat{\beta})$

$= \frac{n}{n-r-1}\hat{\Sigma} = \frac{1}{n-r-1}Y'(I-P_Z)Y.$

Also, $n\hat{\Sigma} = \hat{\varepsilon}'\hat{\varepsilon} \sim W_{m,\,n-r-1}(\Sigma)$ "Wishart Distribution"

Also, $\hat{\Sigma}$ and $\hat{\beta}$ are independent.

under normality assumption,

$\hat{\beta}_{(j)} \sim N_{r+1}(\beta_{(j)}, \sigma_{jj}(Z'Z)^{-1})$

2. 3/29 ~ No class

matlab

$|\hat{\Sigma}| = \det(\hat{\Sigma})$

$e^{\frac{-nm}{2}} = \exp\left(\frac{-nm}{2}\right)$

$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$

Math B7800 3/13/18, Tue

- Likelihood Ratio Test

$$\underset{(r+1)\times m}{\beta} = \begin{bmatrix} \beta_1 \\ \hline \beta_2 \end{bmatrix} \begin{matrix} (q+1)\times m \\ (r-q)\times m \end{matrix}$$

Test: $H_0: \beta_2 = 0$ vs $H_1: \beta_2 \neq 0$.

Partition Z matrix accordingly: $\underset{n\times(r+1)}{Z} = [\underset{n\times(q+1)}{Z_1} \mid \underset{n\times(r-q)}{Z_2}]$

$$\underset{n\times m}{\mathbb{E}[Y]} = Z\beta = [Z_1 \mid Z_2]\begin{bmatrix} \beta_1 \\ \hline \beta_2 \end{bmatrix} = Z_1\beta_1 + Z_2\beta_2$$

"under $H_0: \beta_2 = 0$"

$\mathbb{E}[Y] = Z_1\beta_1$ ($H_0 \Rightarrow$ the predictors in Z_2 are redundant.)

Recall Likelihood Ratio Test.

$$\Lambda = \frac{\max_{H_0} L(\beta,\Sigma)}{\max L(\beta,\Sigma)}$$ "the likelihood ratio"

so, $L(\beta,\Sigma)$ is maximized at "MLE of β,Σ" $\hat{\varepsilon}'$ $\hat{\varepsilon}$

$\beta = \hat{\beta} = (Z'Z)^{-1}Z'Y$ and $\hat{\Sigma} = \frac{1}{n}(Y - Z\hat{\beta})'(Y - Z\hat{\beta})$

and $L(\hat{\beta},\hat{\Sigma}) = (2\pi)^{\frac{-nm}{2}} |\hat{\Sigma}|^{-\frac{n}{2}} e^{\frac{-nm}{2}} = \max L(\beta,\Sigma)$

under H_0, the model becomes $Y = Z_1\beta_1 + \varepsilon$.

so, under H_0, $L(\beta,\Sigma)$ is maximized at

$$\beta = \begin{bmatrix} (Z_1'Z_1)^{-1}Z_1'Y \\ \hline 0 \end{bmatrix} = \begin{bmatrix} \hat{\beta}_1 \\ \hline 0 \end{bmatrix}, \quad \hat{\Sigma}_1 = \frac{1}{n}(Y - Z_1\hat{\beta}_1)'(Y - Z_1\hat{\beta}_1)$$

$$L\left(\begin{bmatrix}\beta_1\\0\end{bmatrix}, \hat{\Sigma}_1\right) = (2\pi)^{\frac{-nm}{2}} |\hat{\Sigma}_1|^{-n/2} e^{\frac{-nm}{2}}$$

so, $\Lambda = \dfrac{L\left(\begin{bmatrix}\beta_1\\0\end{bmatrix}, \hat{\Sigma}_1\right)}{L(\hat{\beta}, \hat{\Sigma})} = \left(\dfrac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}\right)^{n/2}$.

so, $\Lambda^{2/n} = \dfrac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}$, We will reject H_0, when Λ is small or equivalently $\dfrac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}$ is small.

$$U = -\left[n-r-1-\tfrac{1}{2}(m-r+q+1)\right] \ln\left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}\right) \overset{\text{under } H_0}{\sim} \chi^2_{m(r-q)} \quad \text{approximately.}$$

(determinant / determinant)

Reject H_0 at 5% level if $U > \chi_{m(r-q)}$ chi2inv(1-α, m(r-q))

Note: $\dfrac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} = \dfrac{|\hat{\Sigma}|}{|\hat{\Sigma} + (\hat{\Sigma}_1 - \hat{\Sigma})|}$

Since $(\hat{\Sigma}_1 - \hat{\Sigma})$ and $\hat{\Sigma}$ are independent (we proved this only in the univariate case), we need to compare $\hat{\Sigma}$ and $\hat{\Sigma}_1 - \hat{\Sigma}$, in particular, consider the ratio matrix: $(\hat{\Sigma}_1 - \hat{\Sigma})(\hat{\Sigma})^{-1}$.

Let the positive eigenvalues of $(\hat{\Sigma}_1 - \hat{\Sigma})(\hat{\Sigma})^{-1}$ are $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_s > 0$.

- Wilks' lambda: $\prod_{i=1}^{s} \frac{1}{1+\eta_i} = \dfrac{|\hat{\Sigma}|}{|\hat{\Sigma} + (\hat{\Sigma}_1 - \hat{\Sigma})|} = \dfrac{|E|}{|E+H|}$ $\begin{cases} E = \hat{\Sigma} \\ H = \hat{\Sigma}_1 - \hat{\Sigma} \end{cases}$ (1) ~ L_1
- Pillai's trace: $\sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = tr\left((\hat{\Sigma}_1 - \hat{\Sigma})\left((\hat{\Sigma}_1 - \hat{\Sigma}) + \hat{\Sigma}\right)^{-1}\right)$ - (2)
  $= tr\left((\hat{\Sigma}_1 - \hat{\Sigma})(\hat{\Sigma}_1)^{-1}\right)$ ~ L_2
- Hotelling trace: $\sum_{i=1}^{s} \eta_i = tr\left((\hat{\Sigma}_1 - \hat{\Sigma})(\hat{\Sigma})^{-1}\right)$ - (3) ~ L_3
- Roy's greatest root: $\dfrac{\eta_1}{1+\eta_1}$ - (4) ~ L_4

Fact (Linear Algebra).

For two matrices E and H, (both invertible, symmetric, and same size)

$$\frac{|E|}{|E+H|} = \frac{1}{|E^{-1}||E+H|} = \frac{1}{|I+E^{-1}H|}$$

$$= \prod_{i=1}^{s} \frac{1}{1+\eta_i} \quad \text{if the eigenvalues of } E^{-1}H \text{ are } \eta_1, \eta_2, \ldots, \eta_s, 0.$$

(nonzero)

Apply this with $E = \hat{\Sigma}$, $H = \hat{\Sigma}_1 - \hat{\Sigma}$ to obtain (1)

For (2), $\sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = \sum_{i=1}^{s} \frac{1}{1+\frac{1}{\eta_i}}$ ("m-dimension")

$$trace\left(H(E+H)^{-1}\right) = tr\left(H\left[(EH^{-1}+I)H\right]^{-1}\right)$$

$$= tr\left(HH^{-1}(EH^{-1}+I)^{-1}\right) = tr\left((I+EH^{-1})^{-1}\right)$$

If the eigenvalues of HE^{-1} are $\eta_1, \ldots, \eta_s$, then the eigenvalues of EH^{-1} $(HE^{-1})^{-1}$ are $\frac{1}{\eta_1}, \ldots, \frac{1}{\eta_s}$, then the eigenvalues of $(I+EH^{-1})$ are $1+\frac{1}{\eta_1}, \ldots, 1+\frac{1}{\eta_s}$, then the eigenvalues of $(I+EH^{-1})^{-1}$ are $\frac{1}{1+\frac{1}{\eta_1}}, \ldots, \frac{1}{1+\frac{1}{\eta_s}}$

then $tr\left((I+EH^{-1})^{-1}\right) = \sum_{i=1}^{s} \frac{1}{1+\frac{1}{\eta_i}}$ this satisfies (2).

We will reject H_0 if L_1 is small, or L_2 is large, or L_3 is large, or L_4 is large.

Fact (Linear Algebra)

Suppose A and B are two matrices (same size and squared).

A is n.n.d and B is p.d. (so, B is non singular → invertible, $\det(B) \neq 0$.)

Then the eigenvalues of AB^{-1} are non-negative.

(**Fact** For two matrices C and D, the non-zero eigenvalues of CD and DC are same.

using this, all non-zero eigenvalues of AB^{-1} and $\underbrace{B^{-1/2} A B^{-1/2}}_{\text{n.n.d}}$ are same. ("symmetric", n.n.d)

$$\beta = \begin{bmatrix} \beta_{01} & \cdots & \beta_{0m} \\ \beta_{11} & \cdots & \beta_{1m} \\ \vdots & \ddots & \\ \beta_{r1} & \cdots & \beta_{rm} \end{bmatrix}_{(r+1)\times m}$$

$$\hat{\beta} = (Z'Z)^{-1}Z'Y.$$

$$\hat{\beta}' = Y'Z(Z'Z)^{-1}.$$

Prediction

For a future observation with (unknown) response vector Y_0 and (known) predictor values $\underset{\sim}{Z}_0$,

$$\underset{m\text{-entries}}{E[Y_0']} = \underset{\sim}{Z}_0'\beta, \quad \underset{m\times 1}{E[Y_0]} = \underset{m\times(r+1)}{\beta'}\underset{\sim}{Z}_0 \ (r+1)\times 1. \qquad \underset{\sim}{Z}_0 = \begin{bmatrix} 1 \\ Z_{01} \\ \vdots \\ Z_{0r} \end{bmatrix}_{(r+1)+1}$$

→ confidence region ⇒ Prediction region.

1. **Point estimator of both Y_0 and $E(Y_0)$**

$$\hat{\beta}'\underset{\sim}{Z}_0 = Y'Z(Z'Z)^{-1}\underset{\sim}{Z}_0.$$

2. **Confidence region for $E(Y_0)$** ← Y_0 is a vector not a single value. So, not confidence ~~finite~~ interval.

3. **Prediction region for Y_0**

2. $Y_0 \sim N_m(\beta'\underset{\sim}{Z}_0, \Sigma)$, $\hat{\beta}'\underset{\sim}{Z}_0 \sim N_m(\beta'\underset{\sim}{Z}_0, (\underset{\sim}{Z}_0'(Z'Z)^{-1}\underset{\sim}{Z}_0)\Sigma)$

$E[Y_0]$ ← $\begin{bmatrix} Z_0'\beta_{(1)} \\ \vdots \\ Z_0'\beta_{(m)} \end{bmatrix}$

Recall

$\beta = [\beta_{(1)} | \cdots | \beta_{(m)}]$.

$\mathrm{COV}(\beta_{(j)}, \beta_{(k)}) = \sigma_{jk}(Z'Z)^{-1}$

so, $\mathrm{COV}(\underset{\sim}{Z}_0'\beta_{(j)}, \underset{\sim}{Z}_0'\beta_{(k)}) = \underset{\sim}{Z}_0'\sigma_{jk}(Z'Z)^{-1}\underset{\sim}{Z}_0 = (\underset{\sim}{Z}_0'(Z'Z)^{-1}Z_0)\sigma_{jk}$

Hotelling $T^2 = \dfrac{(\hat{\beta}'\underset{\sim}{Z}_0 - \beta'\underset{\sim}{Z}_0)'}{\sqrt{\underset{\sim}{Z}_0'(Z'Z)^{-1}Z_0}} \dfrac{S^{-1}(\hat{\beta}'\underset{\sim}{Z}_0 - \beta'\underset{\sim}{Z}_0)}{\sqrt{\underset{\sim}{Z}_0'(Z'Z)^{-1}\underset{\sim}{Z}_0}}$

, where $S = \dfrac{1}{n-r-1}(Y - Z\hat{\beta})'(Y - Z\hat{\beta})$

$100(1-\alpha)\%$ confidence region $= \left\{\beta'\underset{\sim}{Z}_0 : T^2 \leq \dfrac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha)\right\}$

The $100(1-\alpha)\%$ simultaneous confidence intervals for $\mathbb{E}(Y_i) = \underset{\sim}{z}_0'\underset{\sim}{\beta}_{(i)}$ are

$$\underset{\sim}{z}_0'\hat{\underset{\sim}{\beta}}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m}\right)F_{m,n-r-m}(\alpha)}\sqrt{\underset{\sim}{z}_0'(\underset{\sim}{Z}'\underset{\sim}{Z})^{-1}\underset{\sim}{z}_0\, S_{ii}} \quad , \; i=1,\dots,m$$

where $\hat{\underset{\sim}{\beta}}_{(i)}$ is the ith column of $\hat{\underset{\sim}{\beta}}$ and S_{ii} is the ith diagonal element of S.

Beta_hat(:, i)

3. $Y_0 - \hat{\underset{\sim}{\beta}}'\underset{\sim}{z}_0 = (\underset{\sim}{\beta} - \hat{\underset{\sim}{\beta}})'\underset{\sim}{z}_0 + \underset{\sim}{\varepsilon}_0 \sim N_m(\underset{\sim}{0}, (1 + \underset{\sim}{z}_0'(\underset{\sim}{Z}'\underset{\sim}{Z})^{-1}\underset{\sim}{z}_0)\Sigma)$

The $100(1-\alpha)\%$ prediction ellipsoid for Y_0 becomes

$$(Y_0 - \hat{\underset{\sim}{\beta}}'\underset{\sim}{z}_0)'\, S^{-1}\, (Y_0 - \hat{\underset{\sim}{\beta}}'\underset{\sim}{z}_0) \le (1 + \underset{\sim}{z}_0'(\underset{\sim}{Z}'\underset{\sim}{Z})^{-1}\underset{\sim}{z}_0)\left[\left(\frac{m(n-r-1)}{n-r-m}\right)F_{m,n-r-m}(\alpha)\right]$$

The $100(1-\alpha)\%$ simultaneous prediction intervals for the individual responses Y_{0i}:

$$\underset{\sim}{z}_0'\hat{\underset{\sim}{\beta}}_{(i)} \pm \sqrt{\left(\frac{m(n-r-1)}{n-r-m}\right)F_{m,n-r-m}(\alpha)}\sqrt{(1 + \underset{\sim}{z}_0'(\underset{\sim}{Z}'\underset{\sim}{Z})^{-1}\underset{\sim}{z}_0)\, S_{ii}} \quad , \; i=1,\dots,m.$$

$\underset{\sim}{z}_0'$ Beta_hat(:, i) $S(i,i)$

$\underset{\sim}{Z}_0 = [1 \;\; Z(3, 2{:}\text{end})]'$

so, $Y_3 = \underset{\sim}{z}_0'$ Beta_hat(:, 3).

Math 817800 3/15/18, Thur

Recall Suppose A is pd matrix $(n \times n)$ and symmetric.
(pd: positive-definite)

$$\max\{(\underset{\sim}{b}'\underset{\sim}{x})^2 : \underset{\sim}{x}'A\underset{\sim}{x} \le 1\} = \underset{\sim}{b}'A^{-1}\underset{\sim}{b}.$$

Proof $(\underset{\sim}{b}'\underset{\sim}{x}) = \underbrace{\underset{\sim}{b}'A^{-1/2}}_{\text{vector}}\underbrace{A^{1/2}\underset{\sim}{x}}_{\text{vector}} = (A^{-1/2}\underset{\sim}{b})'(A^{1/2}\underset{\sim}{x})$.

Then

$$(\underset{\sim}{b}'\underset{\sim}{x})^2 \le \left[(A^{-1/2}\underset{\sim}{b})'(A^{-1/2}\underset{\sim}{b})\right]\left[(A^{1/2}\underset{\sim}{x})'(A^{1/2}\underset{\sim}{x})\right]$$
$$= (\underset{\sim}{b}'A^{-1}\underset{\sim}{b})\underbrace{(\underset{\sim}{x}'A\underset{\sim}{x})}_{\le 1} \le \underset{\sim}{b}'A^{-1}\underset{\sim}{b}$$

by Cauchy–Schwartz Inequality

(For two $n \times 1$ vector $\underset{\sim}{x}$ and $\underset{\sim}{y}$, $(\underset{\sim}{x}'\underset{\sim}{y}) \le (\underset{\sim}{x}'\underset{\sim}{x})(\underset{\sim}{y}'\underset{\sim}{y})$
Equality holds if $\underset{\sim}{y} = c\underset{\sim}{x}$ for some constant c.)

Equality holds (if) $A^{1/2}\underset{\sim}{x} = cA^{-1/2}\underset{\sim}{b}$ — ⊛ for some constant c, and c should be such that $\underset{\sim}{x}'A\underset{\sim}{x} = 1$.

⊛ $\underset{\sim}{x} = cA^{-1}\underset{\sim}{b}$ so, $\underset{\sim}{x}'A\underset{\sim}{x} = c^2\underset{\sim}{b}'A^{-1}\underset{\sim}{b} = 1$
$(c\underset{\sim}{b}'A^{-1})A(cA^{-1}\underset{\sim}{b})$ so, $c = \pm\dfrac{1}{\sqrt{\underset{\sim}{b}'A^{-1}\underset{\sim}{b}}}$

The value of $\underset{\sim}{x}$ that attains the max is $\boxed{\pm\dfrac{1}{\sqrt{\underset{\sim}{b}'A^{-1}\underset{\sim}{b}}}A^{-1}\underset{\sim}{b}}$.

Recall the model

$$\underset{n\times m}{Y} = Z\beta + \varepsilon, \quad \underset{n\times m}{Y} = \begin{bmatrix} \underset{\sim}{y}_1' \\ \vdots \\ \underset{\sim}{y}_n' \end{bmatrix}, \quad \underset{n\times(r+1)}{Z} = \begin{bmatrix} \underset{\sim}{z}_1' \\ \vdots \\ \underset{\sim}{z}_n' \end{bmatrix}$$

$$\underset{(r+1)\times m}{\beta} = \left[\beta_{(1)} \mid \cdots \mid \beta_{(m)} \right] \qquad \begin{bmatrix} \beta_{10} & \beta_{20} \\ \beta_{11} & \beta_{21} \\ \beta_{12} & \beta_{22} \end{bmatrix}$$

$$\underset{n\times m}{\varepsilon} = \begin{bmatrix} \underset{\sim}{\varepsilon}_1' \\ \vdots \\ \underset{\sim}{\varepsilon}_n' \end{bmatrix}$$ such that $\underset{\sim}{\varepsilon}_1, \ldots, \underset{\sim}{\varepsilon}_n$ are iid $N_m(\underset{\sim}{0}, \Sigma)$ (assumption) →

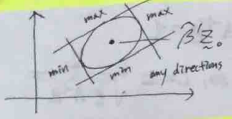
$$\hat{\beta} = (Z'Z)^{-1}Z'Y.$$

Now, we have a future observation with unknown response $\underset{m\times 1}{y_0}$, but known predictors $\underset{(r+1)\times 1}{z_0}$.

$$\underset{m\times 1}{y_0} \sim N\left(\underset{m\times(r+1)}{\beta'} \underset{(r+1)\times 1}{z_0}, \Sigma \right), \quad E[y_0] = \beta' z_0.$$

We saw that 100(1−α)% CR for $\beta' z_0$ is

$$\{ x : (x - \hat{\beta}' z_0)' S^{-1} (x - \hat{\beta}' z_0) \le \frac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha) \}$$



$\beta' z_0$

Q: Obtain simultaneous Scheffe CI for linear combinations of the components of $\beta' z_{0\,(m\times 1)}$.

So, we want CI for $b'\beta' z_0$ for all $m\times 1$ vectors b.

$$\hat{\beta}' z_0 = \{ x : (x - \hat{\beta}' z_0)' A (x - \hat{\beta}' z_0) \le 1 \}, \text{ where}$$

$$A = \frac{S^{-1}}{\frac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha)}$$

If $y = x - \hat{\beta}' z_0$, then the constraint is $y'Ay \le 1$ under this constraint, $(b'y)^2 \le b'A^{-1}b$ by Cachy-Schw. inequality.

So, $\left(b'(x - \hat{\beta}' z_0) \right)^2 \le b' \underbrace{\frac{m(n-r-1)}{n-r-m} S F_{m,n-r-m}(\alpha)}_{\text{"}A^{-1}\text{"}} b$.

So, x represents $\beta' z_0$.

So, the simultaneous CI for $b'\beta' z_0$ is

$$b'\beta' z_0 \pm \sqrt{\frac{m(n-r-1)}{n-r-m} F_{m,n-r-m}(\alpha)\, b'Sb}.$$

$P(\text{the interval } I_b \text{ contains } b'\beta' z_0 \text{ for all } b) = 1-\alpha$.

simultaneous.

• General regression function "joint distribution of Y and $\underset{\sim}{Z}$"

Suppose $(Y, Z_1, \dots, Z_r)$ have a joint distribution with mean

$$\underset{(r+1)\times 1}{\underset{\sim}{\mu}} = \begin{bmatrix} \mu_Y \\ \underset{\sim}{\mu}_Z \end{bmatrix} \quad (\underset{\sim}{\mu}_Z : r\times 1)$$

and covariance

$$\Sigma = \begin{bmatrix} \sigma_{YY} & \underset{\sim}{\sigma}_{YZ}' \\ \underset{\sim}{\sigma}_{YZ} & \Sigma_{ZZ} \end{bmatrix} \quad \begin{matrix} 1\times 1 \\ r\times 1 \end{matrix} \quad \begin{matrix} 1\times r \\ r\times r \end{matrix}$$

Goal: predict Y based on $Z_1, \dots, Z_r$ ← random variables.

We predict Y based on $Z_1, \dots, Z_r$ via some function $g(Z_1, \dots, Z_r)$.

prediction error minimized

prediction error $= Y - g(Z_1, \dots, Z_r)$, $g: \mathbb{R}^r \to \mathbb{R}$.

mean prediction squared error $= \mathbb{E}[(Y - g(Z_1, \dots, Z_r))^2]$

What function minimizes $g(Z_1, \dots, Z_r)$?

Step1 $\mathbb{E}(Y-a)^2 = h(a)$, $h(a)$ is minimized at $a = \mathbb{E}(Y)$.
($\mathbb{E}[Y]$, minimize)

Step2 $\mathbb{E}[(Y-a)^2 | Z] = h(a)$, $h(a)$ is minimized at $\mathbb{E}(Y|Z) = \frac{\mathbb{E}(Y,Z)}{\mathbb{E}(Z)}$

Step3 $\mathbb{E}[(Y-a)^2 | Z_1, \dots, Z_r]$ is minimized at $\mathbb{E}(Y|Z_1, \dots, Z_r) = g^*(Z_1, \dots, Z_r)$.

$$\mathbb{E}[(Y - g(Z_1, \dots, Z_r))^2] = \mathbb{E}\{\underbrace{\mathbb{E}[(Y - g(Z_1, \dots, Z_r))^2 | Z_1, \dots, Z_r]}_{\circledast}\} \quad (*)$$

$$\circledast \quad \mathbb{E}[(Y - g(Z_1, \dots, Z_r))^2 | Z_1, \dots, Z_r] \geq \mathbb{E}[(Y - g^*(Z_1, \dots, Z_r))^2 | Z_1, \dots, Z_r]$$

Taking expectation both sides,

$$\mathbb{E}[(Y - g(Z_1, \dots, Z_r))^2] \geq \mathbb{E}[(Y - g^*(Z_1, \dots, Z_r))^2] \quad \text{by } (*)$$

So, the minimizer $\mathbb{E}(Y | Z_1, \dots, Z_r)$ is called the regression function of Y on $Z_1, \dots, Z_r$. "Jointly normal ⇒ linear regression"

Q: When $\begin{pmatrix} Y \\ Z_1 \\ \vdots \\ Z_r \end{pmatrix} \sim N_{r+1}(\underset{\sim}{\mu}, \Sigma)$, then $\mathbb{E}(Y | Z_1, \dots, Z_r) = ?$

Then the conditional distribution of Y given $Z_1, \dots, Z_r$ is

$$N(\mu_Y + \underset{\sim}{\sigma}_{YZ}' \Sigma_{ZZ}^{-1}(\underset{\sim}{Z} - \underset{\sim}{\mu}_Z), \sigma_{YY} - \underset{\sim}{\sigma}_{YZ}' \Sigma_{ZZ}^{-1} \underset{\sim}{\sigma}_{YZ})$$

So, $\mathbb{E}[Y | Z_1, \dots, Z_r] = \mu_Y + \underset{\sim}{\sigma}_{YZ}' \Sigma_{ZZ}^{-1}(\underset{\sim}{Z} - \underset{\sim}{\mu}_Z)$ ← constant ⇒ linear

$= \beta_0 + \underset{\sim}{\beta}' \underset{\sim}{Z}$, where

$$\beta_0 = (\mu_Y - \underset{\sim}{\sigma}_{YZ}' \Sigma_{ZZ}^{-1} \underset{\sim}{\mu}_Z), \quad \underset{\sim}{\beta} = \Sigma_{ZZ}^{-1} \underset{\sim}{\sigma}_{YZ}$$

• result If the joint distribution is not normal, then $\mathbb{E}(Y | Z_1, \dots, Z_r)$ need not to be linear.

Ex Suppose (Y, Z) has joint pdf, $f_{Y,Z}(Y,Z) = Ze^{-YZ-1}$, $0 \leq Y, Z < \infty$.

Find $\mathbb{E}(Y|Z) = ?$

$$f_{Y|Z}(Y|Z) = \frac{f_{Y,Z}(Y,Z)}{f_Z(Z)} \longrightarrow \text{"marginal"} \quad \frac{Ze^{-YZ-1}}{\int_0^\infty Ze^{-YZ-1}\,dy} = Ze^{-YZ}$$

$$\mathbb{E}[Y|Z] = \int_0^\infty Y \cdot f_{Y|Z}(Y|Z) = \int_0^\infty Y Z e^{-YZ}\,dy = \frac{1}{Z}.$$

So, $\mathbb{E}[Y|Z]$ is not linear in Z.

In general, (without normality) among all linear predictors of y based on $z_1, \dots, z_r$, the one that minimizes the mean squared error is $\beta_0 + \underset{\sim}{\beta}' \underset{\sim}{z}$, where $\beta_0 = (\mu_y - \underset{\sim}{\sigma}_{yz}' \Sigma_{zz}^{-1} \underset{\sim}{\mu}_z)$, $\underset{\sim}{\beta} = \Sigma_{zz}^{-1} \underset{\sim}{\sigma}_{yz}$

Def A linear predictor of y based on $z_1, \dots, z_r$ has the form $b_0 + \underset{\sim}{b}' \underset{\sim}{z}$ for some constants b_0, $\underset{\sim}{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_r \end{pmatrix}$

If $f(b_0, \underset{\sim}{b}) = E\left[(y - b_0 - \underset{\sim}{b}' \underset{\sim}{z})^2\right] \geq f(\beta_0, \underset{\sim}{\beta})$ for any $b_0, \underset{\sim}{b}$.

Math B7800 3/20/18, Tue

• Best Linear Predictor

Recall $(y, z_1, \dots, z_r)$ has a joint distribution with

mean $\underset{\sim}{\mu} = \begin{bmatrix} \mu_y \\ \underset{\sim}{\mu}_z \end{bmatrix}$ ($r+1 \times 1$; 1×1 over $r \times 1$), covariance $\Sigma = \begin{bmatrix} \sigma_{yy} & \underset{\sim}{\sigma}_{yz}' \\ \underset{\sim}{\sigma}_{yz} & \Sigma_{zz} \end{bmatrix}$ ($1 \times r$ over $r \times r$; $(r+1) \times (r+1)$)

Goal: Find a linear predictor of y based on $z_1, \dots, z_r$, which has smallest possible mean squared error.

A linear predictor has the form $b_0 + \underset{\sim}{b}' \underset{\sim}{z}$, $\underset{\sim}{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_r \end{bmatrix}$.

Its mean squared error: $g(b_0, \underset{\sim}{b}) = E\left[(y - b_0 - \underset{\sim}{b}' \underset{\sim}{z})^2\right]$

Goal: Minimize $g(b_0, \underset{\sim}{b})$ w.r.t. $b_0, \underset{\sim}{b}$.

Result: $g(b_0, \underset{\sim}{b}) \geq g(\beta_0, \underset{\sim}{\beta})$, where $\beta_0 = \mu_y - \underset{\sim}{\sigma}_{yz}' \Sigma_{zz}^{-1} \underset{\sim}{\mu}_z$, $\underset{\sim}{\beta} = \Sigma_{zz}^{-1} \underset{\sim}{\sigma}_{yz}$.

proof $g(b_0, \underset{\sim}{b}) = E\left[(y - b_0 - \underset{\sim}{b}' \underset{\sim}{z})^2\right]$ ("$(a+b+c)^2 = a^2+b^2+c^2+2ab+2bc+2ac$")

$= E\left[\left((y - \mu_y) - \underset{\sim}{b}'(\underset{\sim}{z} - \underset{\sim}{\mu}_z) - b_0 + \mu_y - \underset{\sim}{b}' \underset{\sim}{\mu}_z\right)^2\right]$ ($(y - \mu_y - \underset{\sim}{b}' \underset{\sim}{z} + \underset{\sim}{b}' \underset{\sim}{\mu}_z - b_0 + \mu_y - \underset{\sim}{b}' \underset{\sim}{\mu}_z)$)

$= E\left[(y - \mu_y)^2\right] + E\left[\left(\underset{\sim}{b}'(\underset{\sim}{z} - \underset{\sim}{\mu}_z)\right)^2\right] + (b_0 - \mu_y + \underset{\sim}{b}' \underset{\sim}{\mu}_z)^2$

$+ 0 + 0 - 2E\left[(y - \mu_y)\underset{\sim}{b}'(\underset{\sim}{z} - \underset{\sim}{\mu}_z)\right]$

(σ_{yy}; $\underset{\sim}{b}' \Sigma_{zz} \underset{\sim}{b}$; $2E[(y - \mu_y)(\underset{\sim}{b}'(\underset{\sim}{z} - \underset{\sim}{\mu}_z))]$; $2E[(\underset{\sim}{b}'(\underset{\sim}{z} - \underset{\sim}{\mu}_z))(b_0 - \mu_y + \underset{\sim}{b}' \underset{\sim}{\mu}_z)]$)

So, $g(b_0, \underset{\sim}{b}) \geq g(\mu_y - \underset{\sim}{b}' \underset{\sim}{\mu}_z, \underset{\sim}{b})$

$= \sigma_{yy} + \underset{\sim}{b}' \Sigma_{zz} \underset{\sim}{b} - 2\underset{\sim}{b}' \underset{\sim}{\sigma}_{yz}$ — (*)

$$\frac{\partial g(\mu_Y - \underset{\sim}{b}'\underset{\sim}{\mu}_Z, \underset{\sim}{b})}{\partial \underset{\sim}{b}} = 0 + 2\Sigma_{ZZ}\underset{\sim}{b} - 2\underset{\sim}{\sigma}_{YZ}$$

Set $\frac{\partial g}{\partial \underset{\sim}{b}} = 0$, we get $\underset{\sim}{b} = \underbrace{\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}}_{\underset{\sim}{\beta}}$ ← "minimizer of $\underset{\sim}{b}$"

so, $g(b_0, \underset{\sim}{b}) \geq g(\mu_Y - \underset{\sim}{b}'\underset{\sim}{\mu}_Z, \underset{\sim}{b})$

$\geq g(\underbrace{\mu_Y - \underset{\sim}{\beta}'\underset{\sim}{\mu}_Z}_{\beta_0}, \underset{\sim}{\beta}) = g(\beta_0, \underset{\sim}{\beta})$

$\beta_0 = \mu_Y - \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\mu}_Z$, $\underset{\sim}{\beta} = \Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}$.

Smallest mean squared error among linear predictors:

$$\mathbb{E}\left[(Y - \beta_0 - \underset{\sim}{\beta}'\underset{\sim}{Z})^2\right] = \mathbb{E}\left[(Y - (\mu_Y - \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\mu}_Z) - \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{Z})^2\right]$$

$$= \mathbb{E}\left[(Y - \mu_Y + \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\mu}_Z - \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{Z})^2\right]$$

$$= \mathbb{E}\left[((Y - \mu_Y) - \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}(\underset{\sim}{Z} - \underset{\sim}{\mu}_Z))^2\right]$$

$$= g(\beta_0, \underset{\sim}{\beta}) = \sigma_{YY} + \underset{\sim}{\beta}'\Sigma_{ZZ}\underset{\sim}{\beta} - 2\underset{\sim}{\beta}'\underset{\sim}{\sigma}_{YZ}$$

($\underset{\sim}{\beta}' = \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}$, $\underset{\sim}{\beta} = \Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}$)

by plug in ⊛ ↗ $= \sigma_{YY} + \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ} - 2\underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}$

$$= \sigma_{YY} - \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}.$$

$$= \sigma_{YY}\left[1 - \frac{\underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}}{\sigma_{YY}}\right]$$

$$= \sigma_{YY}\left[1 - \rho_{Y(Z)}^2\right], \text{ where } \rho_{Y(Z)} = \pm\sqrt{\frac{\underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}}{\sigma_{YY}}} \text{ is}$$

called the "population multiple correlation coefficient between Y and $\underset{\sim}{Z}$".

Result

$$\max_{b_0, \underset{\sim}{b}} \operatorname{corr}(Y, b_0 + \underset{\sim}{b}'\underset{\sim}{Z}) = \operatorname{corr}(Y, \beta_0 + \underset{\sim}{\beta}'\underset{\sim}{Z}) = \rho_{Y(Z)}$$

Proof

$$\left[\operatorname{corr}(Y, b_0 + \underset{\sim}{b}'\underset{\sim}{Z})\right]^2 = \frac{\left[\operatorname{cov}(Y, b_0 + \underset{\sim}{b}'\underset{\sim}{Z})\right]^2}{\operatorname{var}(Y)\operatorname{var}(b_0 + \underset{\sim}{b}'\underset{\sim}{Z})} = \frac{(\underset{\sim}{b}'\underset{\sim}{\sigma}_{YZ})^2}{\sigma_{YY}\,\underset{\sim}{b}'\Sigma_{ZZ}\underset{\sim}{b}}$$

Recall using Cauchy-Schwartz inequality.

$$(\underset{\sim}{b}'\underset{\sim}{\sigma}_{YZ})^2 = (\underset{\sim}{b}'\Sigma_{ZZ}^{1/2}\Sigma_{ZZ}^{-1/2}\underset{\sim}{\sigma}_{YZ})^2$$

$$\leq (\underset{\sim}{b}'\Sigma_{ZZ}\underset{\sim}{b})(\underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ})$$

so, $\frac{(\underset{\sim}{b}'\underset{\sim}{\sigma}_{YZ})^2}{\underset{\sim}{b}'\Sigma_{ZZ}\underset{\sim}{b}} \leq \underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}$

so, $\left[\operatorname{corr}(Y, b_0 + \underset{\sim}{b}'\underset{\sim}{Z})\right]^2 \leq \frac{\underset{\sim}{\sigma}_{YZ}'\Sigma_{ZZ}^{-1}\underset{\sim}{\sigma}_{YZ}}{\sigma_{YY}}$

"$0 \leq \rho_{Y(Z)}^2 \leq 1$" (Bad) "linearity" between Y, Z (Good)

(If $Y, \underset{\sim}{Z}$ has a joint dist., we can predict Y based on $\underset{\sim}{Z}$, and vice-versa.)

$$= \left(\operatorname{corr}(Y, \beta_0 + \underset{\sim}{\beta}'\underset{\sim}{Z})\right)^2 = \rho_{Y(Z)}^2.$$

Now, suppose we have $\begin{bmatrix} Y \\ \underset{\sim}{Z} \end{bmatrix}$ is jointly $N_{r+1}(\underset{\sim}{\mu}, \Sigma)$, we observe a sample of size n, $\begin{pmatrix} Y_1 \\ \underset{\sim}{Z}_1 \end{pmatrix}, \ldots, \begin{pmatrix} Y_n \\ \underset{\sim}{Z}_n \end{pmatrix}$.

Based on these samples, we find MLE of Best linear predictor, smallest mean squared error, $\rho_{Y(Z)}$.

We compute the sample mean and sample covariance,

$$\hat{\underset{\sim}{\mu}} = \begin{pmatrix} \bar{y} \\ \bar{\underset{\sim}{z}} \end{pmatrix}, \quad S = \left[\begin{array}{c|c} s_{yy} & \underset{\sim}{s}_{yz}' \\ \hline \underset{\sim}{s}_{yz} & S_{zz} \end{array}\right], \quad \text{where } s_{yy} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2.$$

$y'(I_n - P_1)y/(n-1)$ $\qquad (y-\bar{y})'(y-\bar{y})/n-1$

$$(\underset{\sim}{s}_{yz})_k = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(z_{ik}-\bar{z}_k), \quad k=1,\dots,r \qquad (y-\bar{y})'(z-\bar{z})/n-1$$

$$(S_{zz})_{kl} = \frac{1}{n-1}\sum_{i=1}^{n}(z_{ik}-\bar{z}_k)(z_{il}-\bar{z}_l), \quad k,l=1,2,\dots,r$$

$(z-\bar{z})'(z-\bar{z})/n-1$

★ MLE of $\underset{\sim}{\mu}, \Sigma$ are $\hat{\underset{\sim}{\mu}}$, $\frac{n-1}{n}S$.

★ Recall (Invariance property)

If $\hat{\theta}$ is the MLE of θ, then for any function g,

$g(\hat{\theta})$ is an MLE of $g(\theta)$.

Recall ★ $\beta_0 = \mu_y - \underset{\sim}{\Sigma}_{yz}'\Sigma_{zz}^{-1}\underset{\sim}{\mu}_z$, $\underset{\sim}{\beta} = \Sigma_{zz}^{-1}\underset{\sim}{\Sigma}_{yz}$.

So, ★ $\hat{\beta}_0 = \bar{y} - \underset{\sim}{s}_{yz}'S_{zz}^{-1}\bar{\underset{\sim}{z}}$ and $\hat{\underset{\sim}{\beta}} = S_{zz}^{-1}\underset{\sim}{s}_{yz}$

★ MLE for Best Linear predictor

$= \hat{\beta}_0 + \hat{\underset{\sim}{\beta}}'\underset{\sim}{z} = \bar{y} + \underset{\sim}{s}_{yz}'S_{zz}^{-1}(\underset{\sim}{z}-\bar{\underset{\sim}{z}})$ ← $\underset{\sim}{z}$ is a variable "observable"

$$\rho^2_{y(z)} = \frac{\underset{\sim}{\sigma}_{yz}'\Sigma_{zz}^{-1}\underset{\sim}{\sigma}_{yz}}{\sigma_{yy}}$$

★ $$\hat{\rho}^2_{y(z)} = \frac{\underset{\sim}{s}_{yz}'S_{zz}^{-1}\underset{\sim}{s}_{yz}}{s_{yy}}$$

★ MLE for smallest mean squared error

$s_{yy} - \underset{\sim}{s}_{yz}'S_{zz}^{-1}\underset{\sim}{s}_{yz}$ is

$\frac{n-1}{n}\left(s_{yy} - \underset{\sim}{s}_{yz}'S_{zz}^{-1}\underset{\sim}{s}_{yz}\right)$ ★.

● Multiple response case

$\begin{bmatrix} \underset{\sim}{y} \\ \underset{\sim}{z} \end{bmatrix}$ $m\times 1$ / $r\times 1$ $\quad (m+r)\times 1$ have a joint distribution with mean $\underset{\sim}{\mu} = \begin{pmatrix} \underset{\sim}{\mu}_y \\ \underset{\sim}{\mu}_z \end{pmatrix}$ and

$$\operatorname{cov}\begin{pmatrix} \underset{\sim}{y} \\ \underset{\sim}{z} \end{pmatrix} = \underset{\sim}{\Sigma} = \left[\begin{array}{c|c} \underset{m\times m}{\Sigma_{yy}} & \underset{m\times r}{\Sigma_{yz}'} \\ \hline \underset{r\times m}{\Sigma_{yz}} & \underset{r\times r}{\Sigma_{zz}} \end{array}\right]$$

A linear predictor of $\underset{\sim}{y}$ based on $\underset{\sim}{z}$ has the form

$\underset{m\times 1}{\underset{\sim}{b}_0} + \underset{m\times r}{\underset{\sim}{b}'} \underset{r\times 1}{\underset{\sim}{z}}$, where $\underset{\sim}{b}_0$ is a constant vector and b is $m\times r$ matrix.

We minimize mean square and cross products of errors

$\mathbb{E}[(\underset{\sim}{y}-\underset{\sim}{b}_0-\underset{\sim}{b}'\underset{\sim}{z})(\underset{\sim}{y}-\underset{\sim}{b}_0-\underset{\sim}{b}'\underset{\sim}{z})']$ $\quad e\ \underset{\sim}{\varepsilon}\underset{\sim}{\varepsilon}'$ $\qquad \underset{\sim}{A}$ = matrix.

★ The best linear predictor of $\underset{\sim}{y}$ based on $\underset{\sim}{z}$ is

$$\underset{\sim}{\mu}_y - \underset{\sim}{\Sigma}_{yz}'\Sigma_{zz}^{-1}\underset{\sim}{\mu}_z + \underset{\sim}{\Sigma}_{yz}'\Sigma_{zz}^{-1}\underset{\sim}{z} = \underset{\sim}{\beta}_0 + \underset{\sim}{\beta}'\underset{\sim}{z}$$

β

$$= \underset{\sim}{\mu}_y + \underset{\sim}{\Sigma}_{yz}'\Sigma_{zz}^{-1}(\underset{\sim}{z}-\underset{\sim}{\mu}_z)$$

"smallest (in the matrix sense: $\underset{\sim}{A} \geq \underset{\sim}{B}$ for matrices A, B if $\underset{\sim}{A}-\underset{\sim}{B}$ is n.n.d)"

★ mean square and products of errors

$$= \Sigma_{yy} - \underset{\sim}{\Sigma}_{yz}'\Sigma_{zz}^{-1}\underset{\sim}{\Sigma}_{yz} = \underset{\sim}{\Sigma}_{yy\cdot z}$$

$$= \mathbb{E}\left[(y-\beta_0-\underset{\sim}{\beta}'\underset{\sim}{z})(y-\beta_0-\underset{\sim}{\beta}'\underset{\sim}{z})'\right]$$

"Partial correlation coefficient" between Y_k, Y_ℓ eliminating the effect of $\underset{\sim}{Z}$ is the (k,ℓ) corr. coeff. corresponding to the covariance matrix $\underset{\sim}{\Sigma}_{YY\cdot Z}$.

e.g. If $\text{corr}(Y_k, Y_\ell)$ is small (w/ eliminating the effect of $\underset{\sim}{Z}$), Y_k, Y_ℓ are dependent on $\underset{\sim}{Z}$.
& $\text{corr}(Y_k, Y_\ell)$ is large w/ the effect of $\underset{\sim}{Z}$

If $\text{corr}(Y_k, Y_\ell)$ is large w/ eliminating the effect of $\underset{\sim}{Z}$, then Y_k, Y_ℓ are not dependent on $\underset{\sim}{Z}$.
& $\text{corr}(Y_k, Y_\ell)$ is large w/ the effect of $\underset{\sim}{Z}$ they are ~~strictly~~ strongly correlated each other.

$$\underset{\sim}{\Sigma}_{YY\cdot Z} = \begin{bmatrix} \sigma_{Y_1Y_1\cdot Z} & \sigma_{Y_1Y_2\cdot Z} & \cdots & \sigma_{Y_1Y_m\cdot Z} \\ \vdots & & \ddots & \\ \sigma_{Y_mY_1\cdot Z} & & \cdots & \sigma_{Y_mY_m\cdot Z} \end{bmatrix}$$

So, Partial corr. coeff. between Y_k, Y_ℓ eliminating the effect of $\underset{\sim}{Z}$

$$= \frac{\sigma_{Y_kY_\ell\cdot Z}}{\sqrt{\sigma_{Y_kY_k\cdot Z}\cdot\sigma_{Y_\ell Y_\ell\cdot Z}}}$$

Math B7800 3/22/18, Thur

4/10, Tue "Exam 2"
or 4/12, Thur

Chapter 14 of the other text book.

• Non-linear Regression

In linear regression model, $Y = \beta_0 + \beta_1 x + \varepsilon$ (response, predictor, error)

$\mathbb{E}[Y] = \beta_0 + \beta_1 x$ (a linear function both in parameters and predictors)

In nonlinear regression, the relation between $\mathbb{E}[Y]$ and the parameters β_0, β_1, and the predictor x is no longer linear.

"generalized linear model"

An important and useful subclass is the generalized linear model (GLM).
Here, the relation between $\mathbb{E}[Y]$ and β_0, β_1, and x is via a link function η s.t. $\eta(\mathbb{E}[Y]) = \beta_0 + \beta_1 x$.

e.g.
Logistic regression, Probit response function (mean), log-log mean response function.

In many examples, the response variable takes two possible values.

Ex1 While predicting the possibility of heart disease based on ① diet and ② lifestyle measures.

Ex2 While predicting whether a married women will be in the work force based on ① family background, ② education, ③ household income, etc.

Ex3 While predicting whether a home owner will buy liability insurance.

Ex4 While predicting whether a company will have a legal division or not.

Ex5 While deciding whether to give loan/credit card.

Now, we assume Y takes two values: 0 and 1.

Suppose $\pi = \mathbb{P}(Y=1)$, so $\mathbb{P}(Y=0) = 1-\pi$.

$\mathbb{E}[Y] = \pi$, $\text{var}(Y) = \pi(1-\pi)$

If we try to use linear model, $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Issues:

1. $\text{var}(\varepsilon_i) = \text{var}(Y_i) = \underset{\downarrow}{\mathbb{E}[Y_i]}(1-\underset{\downarrow}{\mathbb{E}[Y_i]}) = \mathbb{E}(Y_i)(1-\mathbb{E}(Y_i))$ (ith observation)

(variance depends on mean)

2. Possible values of ε_i: $1-\mathbb{E}(Y_i)$, $0-\mathbb{E}(Y_i)$

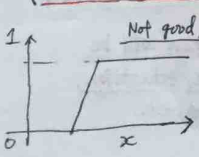
$\Rightarrow 1-\pi_i$, $-\pi_i$.

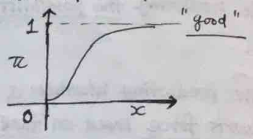
(Normal approximation for ε_i is not good)

*3. $\pi_i = \mathbb{E}(Y_i) = \beta_0 + \beta_1 x_i$ ← $\mathbb{E}(Y_i) = 1\cdot\pi_i + 0\cdot(1-\pi_i) = \pi_i$
$= \mathbb{P}(Y_i = 1)$

Since π_i is a probability, $|\beta_0 + \beta_1 x_i| \le 1$.

(π_i is dependent on x_i)





Ex Y^c = Duration of pregnancy

X = Index for alcohal consumption.

Here, we can use linear model $Y^c = \beta_0^c + \beta_1^c x + \varepsilon^c$, where $\varepsilon \sim N(0, \sigma^2)$

It is considered to be a premature delivery if $Y^c \le 38$ weeks, otherwise full term delivery.

Define

$$Y = \begin{cases} 1 & \text{if } Y^c \le 38 \text{ (preterm)} \\ 0 & \text{if } Y^c > 38 \text{ (full term)} \end{cases}$$

$$\pi = \mathbb{P}[Y=1] = \mathbb{P}[Y^c \le 38] = \mathbb{P}(\beta_0^c + \beta_1^c x + \varepsilon^c \le 38)$$

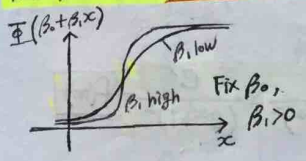
$$= \mathbb{P}\left[\varepsilon_c \le 38 - \beta_0^c - \beta_1^c x\right] = \mathbb{P}\left[\frac{\varepsilon^c}{\sigma} \le \underbrace{\frac{38}{\sigma} - \frac{\beta_0^c}{\sigma}}_{\beta_0} - \underbrace{\frac{\beta_1^c}{\sigma}}_{\beta_1} x\right]$$

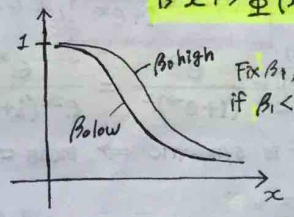
$$= \mathbb{P}[Z \le \beta_0 - \beta_1 x] = \Phi(\beta_0 + \beta_1 x)$$

so, $\Phi^{-1}(\pi) = \beta_0 + \beta_1 x$

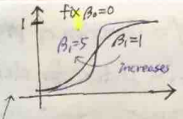
$\Phi^{-1}(\mathbb{E}(Y)) = \beta_0 + \beta_1 x$

This is called "probit mean response function" (probit transformation is $x \mapsto \Phi^{-1}(x)$).

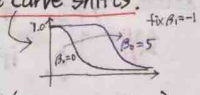




• Properties of the curve



1. Always lie between 0 and 1.
2. Monotone function.
3. For fixed β_0, if β_1 increases, then curve becomes more Sigmoid. (S자모형의)
4. If β_1 is fixed and β_0 is changed, then the curve shifts. (이동하다)



5. If $\mathbb{P}(Y=1)=\pi=\boxed{\Phi(\beta_0+\beta_1 x)}$, then

$$1-\pi=\mathbb{P}(Y=0)=1-\Phi(\beta_0+\beta_1 x)=\boxed{\Phi(-\beta_0-\beta_1 x)}$$

(Using the property $\Phi(z)+\Phi(-z)=1$)

" (read) of meaning."

* This model is hard to interpret value of β_i (parameter)
(probit model)

• Logistic Regression

Logistic distribution is a continuous distribution with pdf

pdf $\quad \boxed{f(x)=\dfrac{e^x}{(1+e^x)^2}},\ x\in\mathbb{R}$

cdf $\quad \boxed{F(x)=\dfrac{e^x}{1+e^x}},\ x\in\mathbb{R}$

$$\boxed{f(-x)}=\frac{e^{-x}}{(1+e^{-x})^2}=\frac{e^x}{e^{2x}(1+e^{-x})^2}=\frac{e^x}{(e^x+1)^2}=\boxed{f(x)}$$

So, f is symmetric $\Rightarrow$ mean $=0$.

$e^{2x}(e^{-2x}+2e^{-x}+1)=(1+2e^x+e^{2x})=(e^x+1)^2$

★ Variance $=\int_{-\infty}^{\infty}x^2\dfrac{e^x}{(1+e^x)^2}dx$ (pdf)

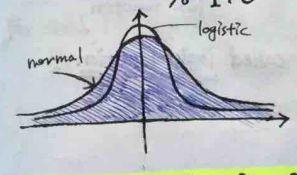
$=2\int_0^{\infty}x^2\dfrac{e^x}{(1+e^x)^2}dx$ ("even function")

f is even iff $f(-x)=f(x)$ for all x in the domain of f.

"by part" $\int u\,dv=uv-\int v\,du$, $u=x^2$, $dv=\dfrac{e^x}{(1+e^x)^2}dx$, $v=-\dfrac{1}{(1+e^x)}$

$$=2\left[\underbrace{x^2\left(-\frac{1}{1+e^x}\right)\Big|_0^{\infty}}_{\Rightarrow 0}+\int_0^{\infty}\frac{2x}{1+e^x}dx\right]$$

$$=4\int_0^{\infty}\frac{x}{1+e^x}dx\overset{\text{check}}{=}4\int_0^{\infty}\log(1+e^{-x})\,dx\overset{\text{check}}{=}\frac{\pi^2}{3}$$



Now, suppose $y^c=\beta_0^c+\beta_1^c x+\varepsilon^c$, where
ε^c has logistic distribution with mean 0, variance σ^2.

$$Y=\begin{cases}1 & \text{if } y^c\le 38\\ 0 & \text{if } y^c>38\end{cases}$$

$$\boxed{\pi}=\mathbb{P}(Y=1)=\mathbb{P}(y^c\le 38)=\mathbb{P}(\varepsilon^c\le 38-\beta_0^c-\beta_1^c x)$$

$$=\mathbb{P}\left(\frac{\varepsilon^c}{\sigma^2}\frac{\pi}{\sqrt{3}}\le\underbrace{\frac{\pi}{\sqrt{3}}38-\frac{\pi}{\sqrt{3}}\beta_0^c}_{\beta_0}-\underbrace{\frac{\pi}{\sqrt{3}}\beta_1^c}_{\beta_1}x\right)$$

($\frac{\varepsilon^c}{\sigma^2}$: "var $=1$")

$$=\boxed{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}\leftarrow \text{cdf}$$

$y = \frac{e^x}{1+e^x}$, then $x = \log\left(\frac{y}{1-y}\right)$

$y + e^x y = e^x$

$\log y + x \log y = x$

$\frac{y}{1-y} = \frac{e^x}{(1+e^x)-e^x} = e^x$

So, $\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x$ (link function)

The function $y \to \log_e\left(\frac{y}{1-y}\right)$ is called logit function.

$\frac{\pi}{1-\pi} = \frac{P(Y=1)}{P(Y=0)}$ ← odd ratio.

Math B7800 — 4/10/18, Tue

"4/26 — Exam II"

~~Tue~~ Thur.

Generalized Linear Model

• Logistic Regression / GLM

we have a binary response y and predictor x.

y takes two values: 0 or 1.

"$Y = E(Y) + \text{error}$"

For the i^{th} observation, $\pi_i = E(y_i) = P(y_i = 1)$.

y_i = response, x_i = predictor

For GLM, $\pi_i = F(\beta_0 + \beta_1 x_i)$, where F is a function (usually F is the cdf of a distribution)

Different choices of F give different GLM. (∫ pdf)

(For a random variable X, the cdf $F_X(x) = P(X \le x)$)

★ If $F(x) = \Phi(x)$, where Φ is the cdf of $N(0,1)$, then we have the probit regression model.

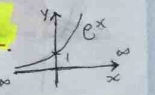
$\pi_i = \Phi(\beta_0 + \beta_1 x_i) \iff \Phi^{-1}(\pi_i) = \beta_0 + \beta_1 x_i$

★ When $F(x) = \frac{e^x}{1+e^x} = 1 - \frac{1}{1+e^x}$ ★ "cdf" of logistic distribution

"increasing"

as $x \to -\infty$, $F(x) = 0$

as $x \to \infty$, $F(x) = 1$



Using this F, we get logistic regression model.

★ $\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} \iff \frac{1}{\pi_i} = 1 + \frac{1}{e^{\beta_0+\beta_1 x_i}} \iff \frac{\pi_i}{1-\pi_i} = e^{\beta_0+\beta_1 x_i}$

$F(\beta_0+\beta_1 x_i)$ //

$\frac{P(y_i=1)}{P(y_i=0)}$ "odds" $\iff \log\frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_i$

$\frac{1}{\pi_i} - 1 = \frac{1-\pi_i}{\pi_i} = \frac{1}{e^{\beta_0+\beta_1 x_i}}$

$\Leftrightarrow \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i$ "odd ratio"

is called odds $= \frac{P(Y_i=1)}{P(Y_i=0)} = \frac{\pi_i}{1-\pi_i}$.

If $F(x) = 1 - e^{-e^x}$ → increasing
is the cdf of Gumbel dist. (as $x \to -\infty$, $F(x)=0$; as $x \to \infty$, $F(x)=1$)

(If $y = 1 - e^{-e^x}$, then $x = \log(-\log(1-y))$)

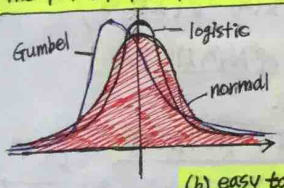
then $F^{-1}(\pi) = \log(-\log(1-\pi))$.

① log-log regression function

Here $\pi_i = 1 - e^{-e^{\beta_0+\beta_1 x_i}} = F(\beta_0+\beta_1 x_i)$

$\Leftrightarrow \log(-\log(1-\pi_i)) = \beta_0 + \beta_1 x_i$.

The pdf of the "three distribution".



Logistic regression is the most popular among the three GLM.
The reason is (a) availability of software for computation;
(b) easy to interpret the parameters.

For logistic regression, $\log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$. odd $= \frac{\pi}{1-\pi} = \frac{P(Y_i=1)}{P(Y_i=0)}$

β_1 captures the change in odds for each unit change in x.

Suppose π and $\tilde{\pi}$ correspond to the predictor value x and $x+1$.

Then $\log\frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$

$\log\frac{\tilde{\pi}}{1-\tilde{\pi}} = \beta_0 + \beta_1 (x+1)$

Subtract $\log\frac{\tilde{\pi}}{1-\tilde{\pi}} - \log\frac{\pi}{1-\pi} = \beta_1$

$= \log\frac{\tilde{\pi}/1-\tilde{\pi}}{\pi/1-\pi}$

☆ Odd Ratio (OR) $= \frac{\tilde{\pi}/(1-\tilde{\pi})}{\pi/(1-\pi)} = e^{\beta_1}$

If $\hat{\beta}_1$ is the MLE of β_1, then $\widehat{OR} = e^{\hat{\beta}_1}$.

☆ How to find MLE of β_0, β_1 in logistic regression.

Need to find the "likelihood function" "joint dist. of y_i"

$L(\beta_0, \beta_1)$ = Joint distribution of the responses.

We have n-individuals with known response and predictor values.

Y	x
Y_1	x_1
⋮	⋮
Y_n	x_n

independent

What is the pmf of Y_1?

$P(Y_1=1) = \pi_1$

$P(Y_1=0) = 1-\pi_1$

$\Leftrightarrow P(Y_1=y) = \pi_1^y (1-\pi_1)^{1-y}$

$g_1(y)$

Then the joint pmf of $Y_1, \dots, Y_n$

$$g(y_1, \dots, y_n) = \prod_{i=1}^{n} g_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

$$\log\left[g(y_1, \dots, y_n)\right] = \sum_{i=1}^{n} \left[y_i \log \pi_i + (1-y_i) \log(1-\pi_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log \frac{\pi_i}{1-\pi_i} + \log(1-\pi_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i (\beta_0 + \beta_1 x_i) - \log\left(1 + e^{\beta_0 + \beta_1 x_i}\right) \right]$$

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

$$1-\pi_i = \frac{1}{1+e^{\beta_0+\beta_1 x_i}}$$

$$\log(1-\pi_i) = -\log\left(1+e^{\beta_0+\beta_1 x_i}\right)$$

$$1 - \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}} = \frac{1+e^{\beta_0+\beta_1 x_i} - e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$$

To obtain MLE $\hat{\beta}_0$ and $\hat{\beta}_1$ of β_0 and β_1, we need to maximize

$$\log L(\beta_0, \beta_1) = \sum_{i=1}^{n} y_i (\beta_0 + \beta_1 x_i) - \log\left(1+e^{\beta_0+\beta_1 x_i}\right)$$

"No closed form expression"

Ex (Effect of programming experience in successfully completing projects in time)

$$Y_i = \begin{cases} 1 & \text{if } i\text{th candidate completed project in time} \\ 0 & \text{otherwise} \end{cases}$$

X_i = amount of programming experience (in months), i is index of person

The dataset is for 25 candidates (available in book) $(i = 1, \dots, 25)$

Logistic regression was used $\hat{\beta}_1 = 0.615$, $\hat{\beta}_0 = -3.05$.

So the fittest logistic regression model:

$$\log \frac{\pi}{1-\pi} = -3.05 + 0.615x = \beta_0 + \beta_1 x$$

In this case, $\widehat{OR} = e^{\hat{\beta}_1} = e^{0.615} = 1.175$.

For two persons, P1, P2, where P2 has one month more experience.

$$\frac{\text{Odds 2}}{\text{Odds 1}} = 1.175 \iff \frac{\text{odds 2} - \text{odds 1}}{\text{Odds 1}} = 0.175 = 17.5\%$$

So for each additional month of experience, the odds for success increase by 17.5%.

(ratio $P(Y_i=1)$ per $P(Y_i=0)$)

I have two future candidates having 10 months and 15 months of experiences.

$\left(\frac{0}{0}\right)$

Relative difference of odds

$$\widehat{OR} = e^{5\hat{\beta}_1}, \quad 5 = |15-10|$$

$$= e^{0.8075} = 2.24 = \frac{\text{Odds 15}}{\text{Odds 10}}$$

$$\frac{\text{Odds 2} - \text{Odds 1}}{\text{Odds 1}} = 1.24$$

$e^{3.075} = \dots$ $\quad \frac{0.615}{5} \cdot 2 = 3.075$ $\quad \frac{0.615}{5} \cdot 2$

Math B7800 4/12/18, Thur

• Logistic Regression

Last time Example showing the interpretation of the parameter β_1, connection with odd Ratio.

So far, we consider the case where one y value is present for each x value. But there can be multiple y values corresponding to one x value.

Ex (coupon distribution)

A company gives out 5 different coupons ($5, $10, $15, $20, $30) to 200 customers for each category.

count how many customers have redeemed those coupons.

($) coupon amount (X)	# customer given	# customers who redeemed (Y)
5	200	55
10	200	74
15	200	95
20	200	122
30	200	144

(Y: sum of Bernoulli → Binomial)

For $i=1,2,\dots,5$, $Y_i \sim Bin(200, \pi_i)$, where $\pi_i = \dfrac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$

So, $\log\dfrac{\pi_i}{1-\pi_i} = \beta_0+\beta_1 x_i = \pi_i'$

pmf of Y_i is $g_i(y_i) = P(Y_i = y_i) = \binom{n}{y_i}\pi_i^{y_i}\cdot(1-\pi_i)^{n-y_i}$ ($n=200$)

$= \binom{200}{y_i}\pi_i^{y_i}\cdot(1-\pi_i)^{200-y_i}$ ← and

$P(Y_i \le y_i) = \sum_{k=0}^{y_i}\binom{n}{k}\pi_i^k(1-\pi_i)^{n-k}$.

cdf.

• **Joint pmf of $Y_1, \dots, Y_5$**

$$g(y_1,\dots,y_5) = \prod_{i=1}^{5} \binom{200}{y_i} \pi_i^{y_i} (1-\pi_i)^{200-y_i}$$

log: $\prod \mapsto \sum$

$$\log g(y_1,\dots,y_5) = \sum_{i=1}^{5} \left[\log\binom{200}{y_i} + y_i \log \pi_i + (200-y_i)\log(1-\pi_i) \right]$$

$+200\log(1-\pi_i) - y_i\log(1-\pi_i)$

$$= \sum_{i=1}^{5} \left[\log\binom{200}{y_i} + y_i \log\frac{\pi_i}{1-\pi_i} + 200\log(1-\pi_i) \right]$$

So, the log likelihood function "maximize this function"

$$\ell(\beta_0,\beta_1) = \log L(\beta_0,\beta_1)$$

($\log\frac{\pi_i}{1-\pi_i}$ = $\beta_0+\beta_1 x_i$; π_i = $e^{\beta_0+\beta_1 x_i}$)

$$= C + \sum_{i=1}^{5} \left[y_i(\beta_0+\beta_1 x_i) - 200\log(1+e^{\beta_0+\beta_1 x_i}) \right]$$

Next, we maximize the above function with respect to β_0, β_1 using numerical method.

The maximizer will be the MLE. find $\hat\beta_0, \hat\beta_1$

• **The Interpretation of β_1**

$$\log\frac{\pi_i}{1-\pi_i} = \beta_0+\beta_1 x_i$$

So, if $\tilde x_i = x_i + 1$ and $\tilde\pi_i, \pi_i$ are the corresponding probabilities.

$$\log\frac{\tilde\pi_i}{1-\tilde\pi_i} - \log\frac{\pi_i}{1-\pi_i} = \beta_1(\tilde x_i - x_i) = \boxed{\beta_1}$$

($\beta_0+\beta_1(x+1)$ and $\beta_0+\beta_1 x$)

$$= \log\frac{\tilde\pi_i/(1-\tilde\pi_i)}{\pi_i/(1-\pi_i)} = \boxed{\log(OR)}$$

($OR = e^{\beta_1}$)

So, $OR = e^{\beta_1}$, $\widehat{OR} = e^{\hat\beta_1}$.

EX Suppose $\hat\beta_1 = 0.01$. What is the % increase in odds if \$12 is offered instead of \$4.

The odds ratio $\widehat{OR} = e^{8\hat\beta_1} = e^{0.08} = 1.083$. ($|12-4| = 8$)

$$\frac{\text{Odds}_{12}}{\text{Odds}_4} = 1.083 \quad \text{So,} \quad \frac{\text{Odds}_{12} - \text{Odds}_4}{\text{Odds}_4} = 0.083$$

$= \frac{\text{Odds}_{12}}{\text{Odds}_4} - 1 = 1.083 - 1$

So, the odds will increase by 8.3% for any x_i.

(π_i depends on x_i; Chance)

(odds does not depend on x_i)

• **Multiple logistic Regression**

Here, we have $p-1$ predictors ($x_1, \dots, x_{p-1}$) and one response (y).

Again, y takes two values: 0 or 1

x_i's are be either categorical or continuous variables.

Ex (Disease occurrence)

$Y = \begin{cases} 1 & \text{if individual infected} \\ 0 & \text{if not} \end{cases}$

Four predictors
- Two continuous (X_1 = Age, X_2 = Cholestrol level)
- Two categorical (X_3, X_4 = Socio economic status, X_5 = city sector)

Socio economic status

	Middle - X_3	X_4 - lower
upper	0	0
Middle	1	0
Lower	0	1

City sector

	X_5 - sector
Sector 1	0
Sector 2	1

So, the predictor $\underset{\sim}{X} = (1, X_1, \dots, X_5)'$

So, we have information about n individuals

we know $(Y_i, \underset{\sim}{x}_i)$, $i = 1, \dots, n$.

The logistic regression model is as follows.

$Y_i \sim \text{Ber}(\pi_i)$, where $\pi_i = \dfrac{e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}{1+e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}}}$, $\underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_5 \end{bmatrix}$

So, the pmf of Y_i is $g_i(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

So, the joint pmf $g(y_1, \dots, y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$

So, $\log g(y_1, \dots, y_n) = \sum_{i=1}^{n}\left[y_i \log\frac{\pi_i}{1-\pi_i} + \log(1-\pi_i)\right]$

$= \sum_{i=1}^{n}\left[y_i(\underset{\sim}{x}_i'\underset{\sim}{\beta}) - \log(1+e^{\underset{\sim}{x}_i'\underset{\sim}{\beta}})\right]$

- Interpretation of the parameters

e^{β_i} is the odd ratio between two cases where ① everything except the i^{th} predictor is kept fixed and ② the i^{th} predictor value is increased by 1.

Suppose $\hat{\beta}_1$ is such that $e^{\hat{\beta}_1} = 1.0309$

this means that the odds for having the disease will go up by 3.09% for each year of age.

Suppose

$e^{\hat{\beta}_5} = 4.82$ (if change "section", odds of disease 5 times. e.g. 5 → 25)

then the odds of having the disease in sector 2 becomes approx. five time the odds in sector 1.

- Inference about the parameters

The inference in this case is for large sample size

The MLE $\hat{\beta}$ approximately follows normal distribution with mean β and covariance $-H(\hat{\beta})$, where

$H_{ij} = \left.\dfrac{\partial^2 \log L(\beta)}{\partial\beta_i\beta_j}\right|_{\hat{\beta}}$ (If $\hat{\beta}$ given, can you find covariance matrix?)

This will allow us to obtain confidence intervals and simultaneous CI for different components of β.

Math B7800 4/17/18, Tue

We are discussing logistic regression.

Ex In a simple logistic regression, $\left(\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x\right)$ if the estimate of β_0 and β_1 are -3 and 1.2, i.e. $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 1.2$ then what is the probability that among 100 future individuals having $x=5$, there will be at least 60 individuals with positive response.

$\mathbb{P}$(an individual having $x=5$ gives positive response)

$$= \frac{e^{\beta_0+5\beta_1}}{1+e^{\beta_0+5\beta_1}} = \pi(5)$$

So, # individuals among the future 100 individuals who will give positive response $- \circledast$

$$\sim \text{Bin}(100, \pi(5))$$

$$\mathbb{P}(\#\circledast \geq 60) = \sum_{k=60}^{100} \binom{100}{k} \pi(5)^k (1-\pi(5))^{100-k}$$

$\widehat{\mathbb{P}(\#\circledast \geq 60)}$ ← plug in $\hat{\beta}_0 = -3$, $\hat{\beta}_1 = 1.2$.

Recall: If $\hat{\theta}$ is an MLE of θ, then for any function g, $g(\hat{\theta})$ is MLE of $g(\theta)$.

So, MLE of the probability is

$$\sum_{k=60}^{100} \binom{100}{k} \left(\hat{\pi}(5)\right)^k \left(1-\hat{\pi}(5)\right)^{100-k}, \text{ where}$$

$$\hat{\pi}(5) = \frac{e^{-3+1.2\times 5}}{1+e^{-3+1.2\times 5}} = \frac{e^3}{1+e^3}$$

M 3/16

Q2: use normal approximation to approximate the estimate.

Recall: $\text{Bin}(n,p) \approx N(np, npq)$, $q = 1-p$.

so the normal approximation is $P(N(100\cdot\hat{\pi}(5), 100\cdot\hat{\pi}(5)(1-\hat{\pi}(5))) \geq 60)$

$$= P(N(95.26, 4.5) \geq 60)$$

$$= 1 - \Phi\left(\frac{60-95.26}{\sqrt{4.5}}\right)$$

($\sqrt{4.5}$ ← standard deviation)

• Inference for the parameters in logistic regression.

the inference is correct for large sample size $\hat{\beta}$ is approximately normal with mean β and covariance matrix $\left[-H(\hat{\beta})\right]^{-1}$ ← this is correct. check the last note.

where $\underset{p\times p}{H(\beta)}$ is the matrix of second derivates

$$H_{ij}(\beta) = \frac{\partial^2 \log L(\beta)}{\partial\beta_i\,\partial\beta_j}, \quad i,j = 0,1,\cdots,p-1$$

Using this approximation,

we obtain CI for β_k, simultaneous CI for the parameters $\beta_0,\cdots,\beta_{k-1}$.

Also, we can test $H_0: \beta_k = 0$ vs $H_1: \beta_k \neq 0$ —(a)

(b) $H_0: \beta_k = 0$ vs $H_1: \beta_k > 0$

(c) $H_0: \beta_k = 1$ vs $H_1: \beta_k \neq 1$.

Q: CI for β_k? (Standard normal $N(0,1)$.)

$$100(1-\alpha)\%\text{ CI for } \beta_k = \hat{\beta}_k \pm z_{\alpha/2}\sqrt{\left(-H(\hat{\beta})\right)^{-1}_{k,k}} \quad -(a)$$

• How to test $H_0: \beta_k = 0$ vs $H_1: \beta_k > 0$

Reject H_0 at level α if $\hat{\beta}_k > z_\alpha \sqrt{\left(-H(\hat{\beta})\right)^{-1}_{k,k}}$ —(b)

(z_α: one side)

• How to test $H_0: \beta_k = 1$ vs $H_1: \beta_k \neq 1$.

Reject H_0 at level α if $|\hat{\beta}_k - 1| > z_{\alpha/2}\sqrt{\left(-H(\hat{\beta})\right)^{-1}_{k,k}}$ —(c)

• $H_0: \beta_k = 1$ vs $H_1: \beta_k < 1$?

Reject H_0 at level α if $\hat{\beta}_k < 1 - z_\alpha \sqrt{\left(-H(\hat{\beta})\right)^{-1}_{k,k}}$.

Likelihood Ratiotest

$q < p$,

$H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0$ vs $H_1 : H_0$ is false

Λ = likelihood ratio $= \dfrac{\sup_{H_0} L(\underset{\sim}{\beta})}{\sup L(\underset{\sim}{\beta})}$

Note that under H_0, we also have a logistic regression model

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{q-1} X_{q-1}$$

using similar approach, we can find MLE of $\beta_0, \ldots, \beta_{q-1}$ under H_0.

Call the MLE $\hat{\underset{\sim}{\beta}}_q$.

So, $\Lambda = \dfrac{L\left(\begin{bmatrix} \hat{\underset{\sim}{\beta}}_q \\ 0 \end{bmatrix}\right)}{L(\hat{\underset{\sim}{\beta}})}$ small → reject H_0.

$-2 \log \Lambda \approx \chi^2_{p-q}$.

reject H_0 if $-2 \log \Lambda > \chi^2_{p-q,\alpha}$ (at level α)

Model selection

Define AIC $= -2 \log L(\hat{\beta}) + 2(\#\text{ variables})$ ← minimize

We compare all possible subsets of predictors, when the total # of predictors is $\leq$ 30 or 40.

If the number of predictors is more than 40, we should use stepwise methods for finding the optional model.

Principal Component Analysis (PCA)

population principal component

suppose $\underset{\sim}{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix}$ is a vector consisting of p features of an individual.

We want to choose linear combinations (any combination of $\underset{\sim}{X}$)

$Y_1 = \underset{\sim}{a}_1' \underset{\sim}{X}$ ($\underset{\sim}{a}_1 = (a_{11}, a_{12}, \ldots a_{1p})$)

$\vdots$

$Y_p = \underset{\sim}{a}_p' \underset{\sim}{X}$ ($\underset{\sim}{a}_p = (a_{p1}, a_{p2}, \ldots, a_{pp})$)

($\underset{\sim}{a}_1, \ldots, \underset{\sim}{a}_p$ are $p \times 1$ vector of constants)

so that $Y_1, \ldots, Y_p$ are uncorrelated and $Y_1, \ldots, Y_p$ have max possible variance.

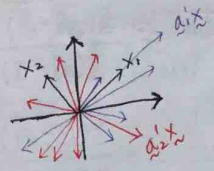
"population"

Suppose Σ is covariance matrix of $\underset{\sim}{X} = [X_1, \ldots, X_p]'$

Then the constant vectors will depend on Σ.

These Y_i's are called (population) principal component of $\underset{\sim}{X}$

Y_1 is the first principal component, Y_2 is the second and so on.



Math B7800 4/19/18, Thur

Question 3 from Exam 1 Let $n=50$, $R^2=0.6$

$Y = X\beta + \varepsilon$, $H_0: \beta_1 = \cdots = \beta_{10} = 0$ vs H_1 : H_0 is false.

$= \beta_0 + \beta_1 X_1 + \cdots + \beta_{10} X_{10} + \varepsilon$

$$Z = \begin{bmatrix} 1 & z_{11} & \cdots & z_{1,10} \\ \vdots & & & \\ 1 & z_{50,1} & \cdots & z_{50,10} \end{bmatrix} = [Z_1 \mid Z_2] \qquad \underset{\sim}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{10} \end{bmatrix} = \begin{bmatrix} \beta_{(1)} \\ \hline \beta_{(2)} \end{bmatrix}$$

(β_0 over $\beta_{(1)}$; $\beta_{(2)} = [\beta_1 \ \cdots \ \beta_{10}]$)

So, $Z_1 = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, $Z_2 = \begin{bmatrix} z_{1,1} & \cdots & z_{1,10} \\ & \ddots & \\ z_{50,1} & \cdots & z_{50,10} \end{bmatrix}$

So, $H_0: \beta_{(2)} = 0$ vs $H_1: \beta_{(2)} \neq 0$.

LRT: $\dfrac{\underset{\sim}{Y}'(P_Z - P_{Z_1})\underset{\sim}{Y}}{\underset{\sim}{Y}'(I - P_Z)\underset{\sim}{Y}}$ ($P_{Z_1} = P_1$) Reject H_0 if LRT is large

$= \dfrac{\underset{\sim}{Y}'(P_Z - P_1)\underset{\sim}{Y}}{\underset{\sim}{Y}'(I - P_Z)\underset{\sim}{Y}} = R^2$ we would reject H_0 at 5% level if $R^2 > \frac{9}{40} F_{9,40}(0.05)$.

Principal Component Analysis (PCA)

Suppose $\underset{\sim}{X}$ is a random vector with mean $\underset{\sim}{0}$ and covariance matrix Σ.

$\underset{\sim}{X}$ represents p features of the individuals in a population. p is usually large.

Goal: Choose fewer (less than p) linear combinations of features without losing much information. → $Y_1, \dots, Y_n$.

$Y_i = \underset{\sim}{a}_i' \underset{\sim}{X}$

Steps Find linear combinations (normalized) $\underset{\sim}{a}_1, \dots, \underset{\sim}{a}_p$ s.t.

$$\underset{\sim}{Y} = \begin{bmatrix} \underset{\sim}{a}_1' \\ \vdots \\ \underset{\sim}{a}_p' \end{bmatrix} \underset{\sim}{X}$$

Y_1

must have uncorrelated components and max possible variance.

$Y_1 = \underset{\sim}{a}_1' \underset{\sim}{X}$

Step 1 • First pricipal component maximize $\underset{\sim}{a}' \Sigma \underset{\sim}{a} = \mathrm{var}(\underset{\sim}{a}' \underset{\sim}{X})$ w.r.t. $\underset{\sim}{a}$ s.t. $\underset{\sim}{a}' \underset{\sim}{a} = 1$.

we saw $\max_{\underset{\sim}{a}:\, \underset{\sim}{a}'\underset{\sim}{a}=1} \underset{\sim}{a}' \Sigma \underset{\sim}{a} = \lambda_1$ ← largest eigen value of Σ

$\max_{\underset{\sim}{a} \neq \underset{\sim}{0}} \frac{\underset{\sim}{a}' \Sigma \underset{\sim}{a}}{\underset{\sim}{a}' \underset{\sim}{a}}$ by spectral decomposition.

$\underset{\sim}{a}_1$ = eigenvector of Σ corresponding to λ_1 achives the max.

Recall: Spectral decomposition of real symmetric matrix.

$A = U \Lambda U'$ for any symmetric real matrix A, where U is orthogonal and Λ is diagonal.
($p \times p$)

The diagonal entries of Λ are the eigenvalues of A and ~~$PP' = I$~~ $UU' = I$.

If $U = [\underset{\sim}{u}_1 | \cdots | \underset{\sim}{u}_p]$ ($p \times p$)

$$U \Lambda U' = [\underset{\sim}{u}_1 | \cdots | \underset{\sim}{u}_p] \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} \begin{bmatrix} \underset{\sim}{u}_1' \\ \hline \\ \underset{\sim}{u}_p' \end{bmatrix}$$

$$= \sum_{i=1}^{p} \lambda_i \underset{\sim}{u}_i \underset{\sim}{u}_i' = A$$

Let $\Sigma = U \Lambda U'$ be the spectral decomposition where

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix}$$ is s.t. $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ ← because Σ is always n.n.d.

Note: $\underset{\sim}{u}_1, \dots, \underset{\sim}{u}_p$ form a basis of $\mathbb{R}^p$.

so, any $\underset{\sim}{a} \in \mathbb{R}^p$ can be expressed as → $\underset{\sim}{a} = x_1 \underset{\sim}{u}_1 + x_2 \underset{\sim}{u}_2 + \cdots + x_p \underset{\sim}{u}_p$ for some constants $x_1, \dots, x_p$.

$\underset{\sim}{a} = \underset{\sim}{x}' \underset{\sim}{u}$

$\underset{\sim}{a} \cdot \underset{\sim}{u}_1 = x_1 \cdot 1 + x_2 \cdot 0 + \cdots + x_p \cdot 0 = x_1$

Similarly, $x_i = (\underset{\sim}{a} \cdot \underset{\sim}{u}_i)$ for $i = 1, \dots, p$.

Then any $\underset{\sim}{a} \in \mathbb{R}^p$ can be written as $\underset{\sim}{a} = \sum_{i=1}^{p} (\underset{\sim}{a} \cdot \underset{\sim}{u}_i) \underset{\sim}{u}_i$

Then $\underset{\sim}{a}'\underset{\sim}{a} = \left(\sum_{i=1}^{p}(\underset{\sim}{a}\cdot\underset{\sim}{u}_i)\underset{\sim}{u}_i'\right)\left(\sum_{j=1}^{p}(\underset{\sim}{a}\cdot\underset{\sim}{u}_j)\underset{\sim}{u}_j\right)$

($\underset{\sim}{x}\underset{\sim}{u}'$, $\underset{\sim}{x}'\underset{\sim}{u}$)

$$= \sum_{i=1}^{p}(\underset{\sim}{a}\cdot\underset{\sim}{u}_i)^2$$

$$\underset{\sim}{a}'\Sigma\underset{\sim}{a} = \left(\sum_{i=1}^{p}(\underset{\sim}{a}\cdot\underset{\sim}{u}_i)\underset{\sim}{u}_i'\right)\left(\sum_{j=1}^{p}\lambda_j\underset{\sim}{u}_j\underset{\sim}{u}_j'\right)\left(\sum_{k=1}^{p}(\underset{\sim}{a}\cdot\underset{\sim}{u}_k)\underset{\sim}{u}_k'\right)$$

$(\underset{\sim}{a}\cdot\underset{\sim}{u}_i)\underset{\sim}{u}_i'\ \lambda_j\underset{\sim}{u}_j\underset{\sim}{u}_j'\ (\underset{\sim}{a}\cdot\underset{\sim}{u}_k)\underset{\sim}{u}_k$

$= \cdots\cdots\ \underset{\sim}{u}_i'\underset{\sim}{u}_j\underset{\sim}{u}_j'\underset{\sim}{u}_k = 0$ if $i\neq j$ or $j\neq k$

$$= \sum_{i=1}^{p}(\underset{\sim}{a}\cdot\underset{\sim}{u}_i)\lambda_i(\underset{\sim}{a}\cdot\underset{\sim}{u}_i) = \sum_{i=1}^{p}\lambda_i(\underset{\sim}{a}\cdot\underset{\sim}{u}_i)^2 \quad -\circledast$$

($(\underset{\sim}{a}\cdot\underset{\sim}{u}_i) = x_i$)

So, what does the problem reduce to?

call $x_i = (\underset{\sim}{a}\cdot\underset{\sim}{u}_i)$

$$\max_{x_1,\cdots,x_p}\frac{\lambda_1x_1^2+\cdots+\lambda_px_p^2}{x_1^2+\cdots+x_p^2} \overset{\circledast}{=} \lambda_1\frac{x_1^2}{x_1^2+\cdots+x_p^2}+\cdots+\lambda_p\frac{x_p^2}{x_1^2+\cdots+x_p^2}$$

(convex combination)

$= \lambda_1$ $\quad = \lambda_1\alpha_1+\cdots+\lambda_p\alpha_p$, where $\alpha_1+\cdots+\alpha_p = 1$.

This shows that $Y_1 = \underset{\sim}{u}_1'\underset{\sim}{X}$ and $\mathrm{var}(Y_1) = \lambda_1$.

(λ_1: of eigen val of Σ)

• Second principal component: $Y_2 = \underset{\sim}{a}_2'\underset{\sim}{X}$ s.t. $\mathrm{cov}(Y_2, Y_1)$

$= \underset{\sim}{a}_2'\Sigma\underset{\sim}{u}_1$ ($\underset{\sim}{u}_1$: ← eigen vector of Σ)

$= \underset{\sim}{a}_2'(\lambda_1\underset{\sim}{u}_1)$ ($\Sigma\underset{\sim}{u}_1 = \lambda_1\underset{\sim}{u}_1$)

$= \lambda_1\underset{\sim}{a}_2'\underset{\sim}{u}_1 = 0$

so, $\underset{\sim}{a}_2'\underset{\sim}{u}_1 = 0$.

maximize $\mathrm{var}(\underset{\sim}{a}_2'\underset{\sim}{X}) = \underset{\sim}{a}_2'\Sigma\underset{\sim}{a}_2$

Now, we need to maximize $\frac{\underset{\sim}{a}_2'\Sigma\underset{\sim}{a}_2}{\underset{\sim}{a}_2'\underset{\sim}{a}_2}$ subject to the constraint $\underset{\sim}{a}_2'\underset{\sim}{u}_1 = 0$

(normalize)

$$\underset{\sim}{a}_2 = \sum_{i=1}^{p}(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_i)\underset{\sim}{u}_i \quad (\underset{\sim}{a}_2\cdot\underset{\sim}{u}_i = x_i)$$

If $\underset{\sim}{a}_2'\underset{\sim}{u}_1 = 0$, then $\underset{\sim}{a}_2 = \sum_{i=2}^{p}(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_i)\underset{\sim}{u}_i$

Also, $\underset{\sim}{a}_2'\Sigma\underset{\sim}{a}_2 = \sum_{i=1}^{p}\lambda_i(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_i)^2 = \sum_{i=2}^{p}\lambda_i(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_i)^2$

$$\frac{\underset{\sim}{a}_2'\Sigma\underset{\sim}{a}_2}{\underset{\sim}{a}_2'\underset{\sim}{a}_2} = \frac{\lambda_2(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_2)^2+\cdots+\lambda_p(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_p)^2}{(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_2)^2+\cdots+(\underset{\sim}{a}_2\cdot\underset{\sim}{u}_p)^2}$$

So, using similar argument,

$$\max_{\substack{\underset{\sim}{a}_2\neq 0\\ \underset{\sim}{a}_2'\underset{\sim}{u}_1=0}}\frac{\underset{\sim}{a}_2'\Sigma\underset{\sim}{a}_2}{\underset{\sim}{a}_2'\underset{\sim}{a}_2} = \lambda_2$$

$\underset{\sim}{a}_2 = \underset{\sim}{u}_2$ (eigen vector) satisfies $\underset{\sim}{a}_2'\underset{\sim}{u}_1 = 0$ and $\frac{\underset{\sim}{a}_2'\Sigma\underset{\sim}{a}_2}{\underset{\sim}{a}_2'\underset{\sim}{a}_2} = \lambda_2$

Thus, $Y_2 = \underset{\sim}{u}_2'\underset{\sim}{X}$, and $\mathrm{var}(Y_2) = \lambda_2$.

($\underset{\sim}{u}_2$: eigen vector of Σ; λ_2: eigen value of Σ)

Following similar argument, $Y_i = \underset{\sim}{u}_i' \underset{\sim}{X}$, $i = 1, \dots, p$

$\mathrm{Var}(Y_i) = \lambda_i$

$\sum_{i=1}^{p} \mathrm{Var}(Y_i) = \sum_{i=1}^{p} \lambda_i = \mathrm{trace}(\Sigma) = \sum_{i=1}^{p} \sigma_{ii}$

($\mathrm{trace}(\Lambda)$)

$\mathrm{tr}(\Sigma) = \mathrm{tr}(P \Lambda P')$, $P'P = I$

$= \mathrm{tr}(\Lambda P'P) = \mathrm{tr}(\Lambda)$.

$= \sum_{i=1}^{p} \underbrace{\mathrm{var}(X_i)}_{\Sigma_{ii}}$ (total variance doesn't change)

- **proportion of total explained by ith principal component (PC)**

$= \dfrac{\lambda_i}{\lambda_1 + \cdots + \lambda_p}$ (this is called portion of variance explained by ith PC)

We will choose first k PC if the corresponding explained variance $\dfrac{\lambda_1 + \cdots + \lambda_k}{\lambda_1 + \cdots + \lambda_p}$ is close to 1.

If $\underset{\sim}{X}$ has mean $\underset{\sim}{\mu}$, covariance Σ, then the (population) PCs are

$Y_1 = \underset{\sim}{u}_1' (\underset{\sim}{X} - \underset{\sim}{\mu}), \dots, Y_p = \underset{\sim}{u}_p' (\underset{\sim}{X} - \underset{\sim}{\mu})$,

where $\underset{\sim}{u}_1, \dots, \underset{\sim}{u}_p$ are the eigenvectors of Σ.

- **PC for standardized random vector.**

Suppose $\underset{\sim}{X}_{p \times 1}$ has mean $\underset{\sim}{0}$, covariance Σ

Define $Z_i = \dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}$, $\underset{\sim}{Z} = \begin{bmatrix} Z_1 \\ \vdots \\ Z_p \end{bmatrix}$

$\mathrm{cov}(Z_i, Z_j) = \mathrm{cov}\left(\dfrac{X_i - \mu_i}{\sqrt{\sigma_{ii}}}, \dfrac{X_j - \mu_j}{\sqrt{\sigma_{jj}}}\right) = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \rho_{ij}$

$\mathbb{E}(\underset{\sim}{Z}) = \underset{\sim}{0}$,

$\mathrm{cov}(\underset{\sim}{Z}) = \rho = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & & & \vdots \\ \vdots & & \ddots & & \\ \rho_{p1} & \rho_{p2} & \cdots & & 1 \end{bmatrix}$

Let $V = \mathrm{Diag}(\Sigma) = \begin{bmatrix} \sigma_{11} & & 0 \\ & \ddots & \\ 0 & & \sigma_{pp} \end{bmatrix}$

$= (V^{1/2})^{-1} \Sigma (V^{1/2})^{-1}$

The PCs of $\underset{\sim}{Z}$ are $\tilde{Y}_1, \dots, \tilde{Y}_p$, where

$\tilde{Y}_i = \underset{\sim}{\tilde{u}}_i \underset{\sim}{Z}$, where $\underset{\sim}{\tilde{u}}_1, \dots, \underset{\sim}{\tilde{u}}_p$ are eigen vectors of ρ

$= \underset{\sim}{\tilde{u}}_i (V^{1/2})^{-1} (\underset{\sim}{X} - \underset{\sim}{\mu})$

Math B7800 04224 4/24/18, Tue

Principal Components

Last time population PC for standardized covariance.

• special covariance matrices

case1: when Σ is diagonal, $\Sigma = \begin{bmatrix} \sigma_{11} & & 0 \\ & \ddots & \\ 0 & & \sigma_{pp} \end{bmatrix}$

Here, the eigenvalues of Σ are $\sigma_{11}, \dots, \sigma_{pp}$, we order them in decreasing order $\sigma_{(1)} \geq \cdots \geq \sigma_{(p)}$.

$\underset{\sim}{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}$

- First principal component $= \left(X_{(1)} - \mu_{(1)}\right)$
- The ith PC is $\left(X_{(i)} - \mu_{(i)}\right)$

Ex If $\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 3 \end{bmatrix}$ and $\underset{\sim}{X}$ has mean $\underset{\sim}{\mu}$ and cov Σ.

then $PC1 = X_3 - \mu_3$

$PC2 = X_1 - \mu_1$

$PC3 = X_2 - \mu_2$

corresponding correlation matrix $= I$. ← all eigenvalues are equal important.

Case 2: All pairwise correlations of the components are equal to ρ.

Suppose $\text{var}(X_i) = \sigma_{ii}$.

Then $\text{cov}(X_i, X_j) = \begin{cases} \rho\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}} & \text{if } i \neq j \\ \sigma_{ii} & \text{if } i = j \end{cases}$ (no correlation.)

so,

$$\Sigma = \begin{bmatrix} \sigma_{11} & \rho\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \cdots & \rho\sqrt{\sigma_{11}}\sqrt{\sigma_{pp}} \\ \vdots & \ddots & & \vdots \\ \rho\sqrt{\sigma_{pp}}\sqrt{\sigma_{11}} & \cdots & \cdots & \sigma_{pp} \end{bmatrix}$$

corresponding correlation matrix $= \begin{bmatrix} 1 & \rho & \rho & \cdots & \rho \\ \rho & 1 & & & \rho \\ \vdots & & \ddots & & \\ \rho & \rho & \cdots & & 1 \end{bmatrix} = \rho$

Step 1: $\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n \times n} = \underset{\sim}{1}_n \underset{\sim}{1}_n'$, $\underset{\sim}{1}_n$ is $n \times 1$ vector of all 1

$\underset{\sim}{1}_n' \underset{\sim}{1}_n = n$

the eigenvalues of $\underset{\sim}{1}_n \underset{\sim}{1}_n'$ are $(n, 0, 0, \ldots, 0)$

$(\underset{\sim}{1}_n \underset{\sim}{1}_n') \cdot \underset{\sim}{1}_n = n \cdot \underset{\sim}{1}_n$. So, $\underset{\sim}{1}_n$ is an eigenvector of $\underset{\sim}{1}_n \underset{\sim}{1}_n'$ corresponding to eigenvalue n.

If $\underset{\sim}{u}$ is an eigenvector of $\underset{\sim}{1}_n \underset{\sim}{1}_n'$ corresponding to eigenvalue 0, then

$(\underset{\sim}{1}_n \underset{\sim}{1}_n')\underset{\sim}{u} = \underset{\sim}{0} = \underset{\sim}{1}_n(\underset{\sim}{1}_n'\underset{\sim}{u}) = (\underset{\sim}{1}_n'\underset{\sim}{u})\underset{\sim}{1}_n$. So, $\underset{\sim}{1}_n'\underset{\sim}{u} = 0$

(where $\underset{\sim}{1}_n'\underset{\sim}{u} = 0$)

Def the eigenspace of a matrix A corresponding to eigenvalue λ is $\{\underset{\sim}{x} : A\underset{\sim}{x} = \lambda\underset{\sim}{x}\}$.

In our case, the eigenspace of $\underset{\sim}{1}_n \underset{\sim}{1}_n'$ (A) corresponding to eigenvalue is $\text{Null}(\underset{\sim}{1}_n')$

$\dim(N(A)) = \dim \text{col}(A) - \text{rank}$.

(n, 1, A)

In this case, $\dim(\text{eigenspace}) = n - 1$.

Any basis of $N(\underset{\sim}{1}_n')$, or in particular, any set of mutually orthogonal vectors $\underset{\sim}{u}_1, \ldots, \underset{\sim}{u}_{n-1}$ satisfying $\underset{\sim}{1}_n'\underset{\sim}{u}_i = 0$, $i = 1, \ldots, n-1$. (basis) and be used as $PC_2, \ldots, PC_n$ for $\underset{\sim}{1}_n \underset{\sim}{1}_n'$.

$\left[\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, \ldots, 0\right]$ "normalized" "mutually orthogonal" (orthogonal.)

$\left[\frac{1}{\sqrt{2\times3}}, \frac{-1}{\sqrt{2\times3}}, \frac{-2}{\sqrt{2\times3}}, 0, \ldots, 0\right]$

$\vdots$

$\left[\underbrace{\frac{1}{\sqrt{(i-1)i}}, \frac{1}{\sqrt{(i-1)i}}, \ldots, \frac{1}{\sqrt{(i-1)i}}}_{(i-1)\text{ times}}, \frac{-(i-1)}{\sqrt{(i-1)i}}, 0, 0, \ldots, 0\right] = \underset{\sim}{e}_i$ for $i = 2, 3, \ldots, n$.

$\frac{(i-1)+(i-1)^2}{(i-1)i} = 1 = \text{norm}$

then $\underset{\sim}{e}_1 = \underset{\sim}{1}_n, \underset{\sim}{e}_2, \ldots, \underset{\sim}{e}_n$ are independent eigenvectors of $\underset{\sim}{1}_n \underset{\sim}{1}_n'$.

In the example, the correlation matrix was

$$\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho \\ \rho & \cdots & \rho & 1 \end{bmatrix}_{p\times p} = \rho \underset{\sim}{1}_p \underset{\sim}{1}_p' + (1-\rho) I_p$$

If $\lambda_1, \ldots, \lambda_p$ are the eigenvalues of $A_{p\times p}$ with corresponding eigenvectors $\underset{\sim}{u}_1, \ldots, \underset{\sim}{u}_p$, then what are the eigenvalues and eigenvectors of $aI_p + bA$?

$a+b\lambda_1, a+b\lambda_2, \ldots, a+b\lambda_p$ are the eigenvalues with eigen vectors $\underset{\sim}{u}_1, \ldots, \underset{\sim}{u}_p$.

$$A\underset{\sim}{u}_i = \lambda_i \underset{\sim}{u}_i$$

$$bA\underset{\sim}{u}_i = b\lambda_i \underset{\sim}{u}_i$$

$$a\underset{\sim}{u}_i + bA\underset{\sim}{u}_i = a\underset{\sim}{u}_i + b\lambda_i \underset{\sim}{u}_i = (a+b\lambda_i)\underset{\sim}{u}_i$$

$$\| \quad (aI + bA)\underset{\sim}{u}_i$$

So, the eigen values of $\rho \underset{\sim}{1}_p \underset{\sim}{1}_p' + (1-\rho) I_p$ are

$$\rho p + (1-\rho),\ 1-\rho,\ 1-\rho,\ \ldots,\ 1-\rho.$$

$$\| \quad 1 + (p-1)\rho$$

so, $PC_1 = \left(\frac{1}{\sqrt{p}}, \frac{1}{\sqrt{p}}, \ldots, \frac{1}{\sqrt{p}}\right)' (\underset{\sim}{x} - \underset{\sim}{\mu})$

$$\vdots$$

$$PC_i = \left(\frac{1}{\sqrt{(i-1)i}}, \frac{1}{\sqrt{(i-1)i}}, \ldots, \frac{1}{\sqrt{(i-1)i}}, \frac{-(i-1)}{\sqrt{(i-1)i}}, 0, \ldots, 0\right)' (\underset{\sim}{x} - \underset{\sim}{\mu})$$

for $i = 2, \ldots, p$.

proportion of variance explained by the first component.

total variance : sum of eigenvalues.

total variance = p.

so, the proportion $= \frac{1+(p-1)\rho}{p} = \rho + \frac{1-\rho}{p}$

If ρ is large and/or p is large, then this propotion is quite large.

If $\rho = 0.8$, $p = 5$, then the proportion is 0.84.

so, it is enough to consider the first component.

● Sample PC (Principal Components)

Suppose $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ are n iid sample from a population with mean $\underset{\sim}{\mu}_{p\times 1}$ and covariance $\Sigma_{p\times p}$.

Obtain sample mean $\bar{x}$ and sample cov $S = \frac{1}{n-1}\sum_{i=1}^{n} (\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})'$.

then obtain eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ of S with corresponding eigenvectors $\hat{\underset{\sim}{e}}_1, \ldots, \hat{\underset{\sim}{e}}_p$.

then the ith sample PC is $\hat{\underset{\sim}{e}}_i' (\underset{\sim}{x} - \bar{\underset{\sim}{x}})$

If S is the sample covariance matrix, then for any constant vector $\underset{\sim}{a}$, what is the sample variance of $\underset{\sim}{a}'\underset{\sim}{x}_1, \ldots, \underset{\sim}{a}'\underset{\sim}{x}_n$?

Ans $\underset{\sim}{a}' S \underset{\sim}{a}$

For two constant vectors $\underset{\sim}{a}, \underset{\sim}{b}$, sample cov between $\underset{\sim}{a}'\underset{\sim}{x}, \underset{\sim}{b}'\underset{\sim}{x}$ is $\underset{\sim}{a}' S \underset{\sim}{b}$.

keeping these in mind, the sample PC maximizes sample variance subject to uncorrelated components.

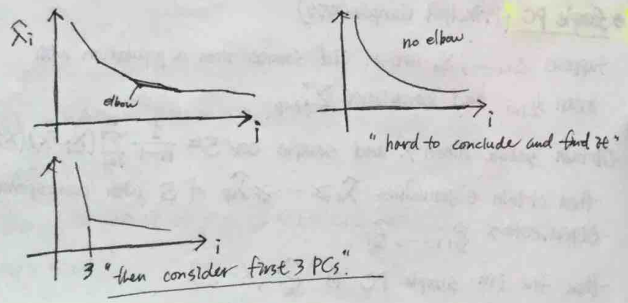
So, we want to find linear combinations $\hat{e}_1, \dots, \hat{e}_p$ s.t.

if $Y_i = \hat{e}_i'(X - \bar{X})$, then the components $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_p \end{pmatrix}$

have zero sample covariance and max sample variance.

Number of PC

Look at the scree plot $(i, \hat{\lambda}_i)$ and look for 'elbow'.



"hard to conclude and find it"

"then consider first 3 PCs".

Math B7800 5/1/18, Tue

Final exam is similar with Exam 1. with 4 page sheets. calculator.
(5/22)

Sample PC

$X_1, X_2, \dots, X_n$ are iid $p \times 1$ random vectors from a population with mean μ and over Σ.

Obtain sample mean $\bar{X}$ and sample covariance matrix S.

Obtain eigenvalues, eigenvectors $(\hat{\lambda}_i, \hat{e}_i)$ $i = 1, \dots, p$ for the matrix S.

then the sample PC are given by $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$.

$\hat{e}_i'(X - \bar{X})$, $i = 1, \dots, p$.

$\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$

Interpretation

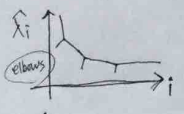
Total sample variance = trace(S)

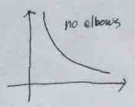
$$= \sum_{i=1}^{p} \hat{\lambda}_i$$

$$= \sum_{i=1}^{p} \underbrace{\text{Sample variance of } \hat{e}_i'(X - \mu)}_{= \hat{\lambda}_i}$$

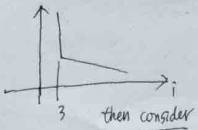
The PC corresponding to small $\hat{\lambda}_i$ are dropped.

Either use Scree plot or a thumb rule applied to standardized sample PC.





"hard to conclude and find it"



then consider first 3 PCs.

Standardized sample PC

Initially we have $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$.

Sample mean $\bar{\underset{\sim}{X}}$, sample cov $= S$, $D = \text{Diag}(S)$.

$\underset{\sim}{Z}_i = D^{-1/2}(\underset{\sim}{X}_i - \bar{\underset{\sim}{X}})$ standardized $\underset{\sim}{X}_i$

$\underset{\sim}{Z}_1, \ldots, \underset{\sim}{Z}_n$ are standardized version of $\underset{\sim}{X}_i$'s.

$\bar{\underset{\sim}{Z}} = \underset{\sim}{0}$ and $\frac{1}{n-1}\sum_{i=1}^{n} \underset{\sim}{Z}_i \underset{\sim}{Z}_i' =$ sample cov matrix of $\underset{\sim}{Z}_1, \ldots, \underset{\sim}{Z}_n$.

$= R$ (sample correlation matrix for $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$)

$P \times P$

trace $(R) = P$, If $\hat{\hat{\lambda}}_i$, $i = 1, \ldots, P$ are the eigenvalues of R.

then $\sum_{i=1}^{P} \hat{\hat{\lambda}}_i = P$

The PC for which $\hat{\hat{\lambda}} \leq 1$ are dropped.

Standardized sample PC are

$\hat{\hat{e}}_i' Z = \hat{\hat{e}}_i' D^{-1/2}(\underset{\sim}{X} - \bar{\underset{\sim}{X}})$, where $\hat{\hat{e}}_i$ is an eigenvector corresponding to $\hat{\hat{\lambda}}_i$ and $\hat{\hat{e}}_1, \ldots, \hat{\hat{e}}_P$ are orthogonal.

(3) The distribution of $\hat{\lambda}_i$ and $\hat{e}_i$, $i = 1, \ldots, P$ are asymptotically independent.

Ex Suppose the spectral decomposition of $S_{3\times 3}$

$\hat{\lambda}_1 = 10$ $\quad \hat{\underset{\sim}{e}}_1 = \begin{bmatrix} 1/\sqrt{3} \\ 1/\sqrt{3} \\ 1/\sqrt{3} \end{bmatrix}$, $\hat{\underset{\sim}{e}}_2 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{bmatrix}$, $\hat{\underset{\sim}{e}}_3 = \begin{bmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ -2/\sqrt{6} \end{bmatrix}$

$\hat{\lambda}_2 = 5$

$\hat{\lambda}_3 = 1$

Q: Find 95% CI for $\lambda_1, \lambda_2, \lambda_3$. ($n$ is large), $n = 80$.

For λ_1, $-1.96 \leq \dfrac{\sqrt{80}(10 - \lambda_1)}{\sqrt{2 \cdot 10^2}} \leq \underbrace{1.96}_{Z_{\alpha/2}}$ $\quad \alpha = 0.05$.

$\Downarrow$

$? \leq \lambda_1 \leq ?$

Sample PC estimate population PC.

Recall $Y_i = e_i'(X-\mu)$, $i=1,\dots,p$ where the population PC.

Recall $\mathrm{cov}(X) = \Sigma$, Σ has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$, with corresponding eigenvectors $e_1, \dots, e_p$.

$$\mathrm{cov}(Y) = \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} = \Lambda \quad , \quad \mathrm{cov}(Y_i, Y_j) = e_i' \Sigma e_j = \begin{cases} 0 & \text{if } i \neq j \\ \lambda_i & \text{if } i = j \end{cases}$$

large sample inference — find confidence interval

(1) $$\sqrt{n}\left[\begin{pmatrix}\hat\lambda_1 \\ \vdots \\ \hat\lambda_p\end{pmatrix} - \begin{pmatrix}\lambda_1 \\ \vdots \\ \lambda_p\end{pmatrix}\right] \simeq N_p(0, 2\Lambda^2)$$

(2) $\sqrt{n}\,(\underset{p\times 1}{\hat e_k} - e_k) \simeq N_p(0, E_k)$, where

$XX' \leftarrow$ rank is 1.

$$E_k = \sum_{\substack{i=1 \\ i\neq k}}^{p} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} e_i e_i' \quad \text{(rank 1)} \quad \leftarrow \text{rank } p-1 \Rightarrow \text{singular matrix}$$

Note that rank (E_k) is $p-1$.

This is fine because $\hat e_k$ satisfies $\sum_{i=1}^{p} (\hat e_{ki})^2 = 1$, and so the vector $\hat e_k$ lies on $p-1$ dimensional sphere.

Q: Find 99% CI for e_{21}.

$$-z_{\alpha/2} \leq \frac{\sqrt{80}\,(1/\sqrt{2} - e_k)}{\sqrt{(\hat E_2)_{1,1}}} \leq z_{\alpha/2} \qquad \alpha = 0.01$$

$$E_2 = \sum_{\substack{i=1 \\ i\neq 2}}^{3} \frac{\lambda_i \lambda_k}{(\lambda_i - \lambda_k)^2} e_i e_i' = \frac{\lambda_1 \lambda_2}{(\lambda_1-\lambda_2)^2} e_1 e_1' + \frac{\lambda_2 \lambda_3}{(\lambda_2-\lambda_3)^2} e_3 e_3'$$

$$(\hat E_2)_{1,1} = \frac{10 \cdot 5}{5^2}\frac{1}{3} + \frac{5}{16}\frac{1}{6}$$

Q: Find 95% CR for $\begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix}$. A ← $(\hat E_2)_{1:2,1:2}$.

$$\sqrt{80}\begin{pmatrix} \hat e_{21} - e_{21} \\ \hat e_{22} - e_{22} \end{pmatrix} \sim N_2\left(0, \begin{bmatrix} a & b \\ b & c \end{bmatrix}\right)$$

$$a = (\hat E_2)_{1,1} \qquad \hat E_2 = \frac{10\times 5}{5^2}\hat e_1 \hat e_1' + \frac{5\times 1}{16}\hat e_3 \hat e_3'$$

$$\sqrt{80}\, A^{-1/2}\begin{pmatrix} \hat e_{21} - e_{21} \\ \hat e_{22} - e_{22} \end{pmatrix} \sim N_2(0, I_2)$$

$$80\begin{pmatrix} \hat e_{21} - e_{21} \\ \hat e_{22} - e_{22} \end{pmatrix}' A^{-1}\begin{pmatrix} \hat e_{21} - e_{21} \\ \hat e_{22} - e_{22} \end{pmatrix} \approx \chi^2_2$$

$$CR := \left\{\begin{pmatrix} e_{21} \\ e_{22} \end{pmatrix} : T \leq \chi^2_{2,0.05}\right\}$$

Test $\underset{\sim}{X}_1, \cdots, \underset{\sim}{X}_n$ iid with mean μ and cov Σ and correlation matrix ρ.

H_0: All pairwise correlations are same $\rho = \rho_0 = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & & \\ \vdots & & \ddots & \\ \rho & & \cdots & 1 \end{bmatrix}$

H_1: H_0 is false.

- Likelihood Ratio test at level α:

First obtain sample correlation matrix R, $R_{ij} = \dfrac{S_{ij}}{\sqrt{S_{ii}}\sqrt{S_{jj}}}$

$\bar{r}_k$ = average of the off-diagonal entries in the k^{th} row of R $\quad k = 1, \cdots, p$.

$$\bar{r}_k = \frac{1}{p-1} \sum_{\substack{i=1 \\ i \neq k}}^{p} R_{ik}$$

$$\bar{r} = \text{average of all off diagonal} = \frac{1}{p(p-1)} \sum\sum_{i \neq k} R_{i,k}$$

$$\text{Gamma } \Gamma = \frac{(p-1)(1-(1-\bar{r})^2)}{p-(p-2)(1-\bar{r})^2},$$

$$T = \frac{p-1}{(1-\bar{r})^2} \left[\sum\sum_{i \neq k} (\bar{r}_{ik} - \bar{r})^2 - \Gamma \sum_k (\bar{r}_k - \bar{r})^2 \right]$$

(Γ: Gamma)

$$\underset{H_0}{\sim} \chi^2_{(p+1)(p-2)/2}$$

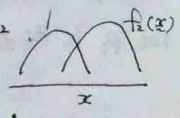
Math B7800 5/3/18/ Thur

Ch.11 Classification

In this set up, we suppose that the data (observation vector, feature vector) comes from one of the classes (or groups).
The class labels of the data vectors are unknown.
The set of possible class is known.
We would like to predict the class labels.

Two classes Suppose π_1 and π_2 denote the two possible classes that the data vectors $\underset{\sim}{x}$ can come from.

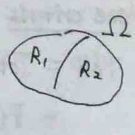
The pdf of $\underset{\sim}{X}$ is $\begin{cases} f_1(\underset{\sim}{x}) & \text{if it is from } \pi_1 \\ f_2(\underset{\sim}{x}) & \text{if it is from } \pi_2 \end{cases}$



Ω = set of all possible values of $\underset{\sim}{x}$.

We want to divide Ω into two disjoint regions R_1 and R_2

so that $\underset{\sim}{x} \in R_i$ will be classified to π_i.



Misclassification Probability

$P(2|1) = P(\text{observation from } \pi_1 \text{ is classified to } \pi_2)$

$= P(\text{observation classied to } \pi_2 \mid \text{it came from } \pi_1)$

$= \int_{R_2} f_1(\underset{\sim}{x}) d\underset{\sim}{x}$

Similarly,

$P(1|2) = \int_{R_1} f_2(\underset{\sim}{x}) d\underset{\sim}{x}$

In many cases, we have prior knowledge about the probability that $\underset{\sim}{X}$ comes from π_1 or π_2.

Let p_i be the prior prob. of π_i, i.e., $P(\underset{\sim}{X} \in \pi_i) = p_i$

$p_1 + p_2 = 1$

- One criteria to minimize

TMP = Total misclassification prob.

$= P(\underset{\sim}{X} \text{ is misclassfied})$

$= P(2|1)\, p_1 + P(1|2)\, p_2$

$= p_1 \int_{R_2} f_1(\underset{\sim}{x}) d\underset{\sim}{x} + p_2 \int_{R_1} f_2(\underset{\sim}{x}) d\underset{\sim}{x}$

~~Cost~~

- Cost involved in Misclassification

	π_1	π_2
π_1	0	$C(1\|2)$
π_2	$C(2\|1)$	0

"Actual"

Your classification

- Average of Expected cost of Misclassification

ECM = E(cost)

$= p_1 P(2|1) C(2|1) + p_2 P(1|2) C(1|2)$

$= p_1 \int_{R_2} f_1(\underset{\sim}{x}) d\underset{\sim}{x}\, C(2|1) + p_2 \int_{R_1} f_2(\underset{\sim}{x}) d\underset{\sim}{x}\, C(1|2)$

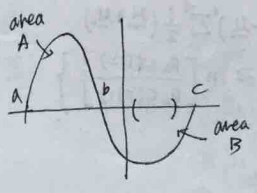
Assume that we know $p_1, p_2, C(1|2), C(2|1), f_1(\underset{\sim}{x}), f_2(\underset{\sim}{x})$.

Goal: find R_1, R_2 so that ECM is minimized.

$\int_{R_1} f_1(\underset{\sim}{x}) d\underset{\sim}{x} + \int_{R_2} f_1(\underset{\sim}{x}) d\underset{\sim}{x} = 1$

so, ECM $= p_1 \left(1 - \int_{R_1} f_1(\underset{\sim}{x}) d\underset{\sim}{x}\right) C(2|1) + p_2 \int_{R_1} f_2(\underset{\sim}{x}) d\underset{\sim}{x}\, C(1|2)$

$= \int_{R_1} \left[p_2 C(1|2) f_2(\underset{\sim}{x}) - p_1 C(2|1) f_1(\underset{\sim}{x}) \right] d\underset{\sim}{x} + p_1 C(2|1)$



$\int_R h(x)dx$ is minimum or max when $R = [b, c]$ when $R = [a, b]$.

$-B \le \int_R f(\underset{\sim}{x}) d\underset{\sim}{x} \le A$

So, using similar argument, ECM will be minimized

$$\text{if } R_1 = \left\{ \underset{\sim}{x} : \underbrace{\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})}}_{\text{ratio of pdf}} \geq \underbrace{\frac{c(1|2)}{c(2|1)}}_{\text{ratio of cost}} \underbrace{\frac{p_2}{p_1}}_{\text{ratio of prior}} \right\}.$$

Suppose π_1 is $N_p(\underset{\sim}{\mu}_1, \Sigma)$ where $\underset{\sim}{\mu}_1 \neq \underset{\sim}{\mu}_2$

and π_2 is $N_p(\underset{\sim}{\mu}_2, \Sigma)$ but same covariance.

Assume $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \Sigma$ are known cov.

Then $f_1(\underset{\sim}{x}) = \frac{1}{(2\pi)^{1/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)\right]$

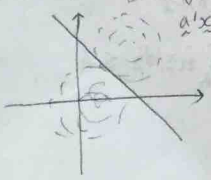
$$f_2(\underset{\sim}{x}) = \frac{1}{(2\pi)^{1/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right]$$

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = \exp\left[-\frac{1}{2}\left\{(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1) - (\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right\}\right]$$

$$= \exp\left[(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\left(\underset{\sim}{x} - \frac{\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2}{2}\right)\right]$$

so, $R_1 = \left\{ \underset{\sim}{x} : \underbrace{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}}_{a'}\underset{\sim}{x} - (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\frac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) \geq \ln\left[\frac{p_2\, c(1|2)}{p_1\, c(2|1)}\right] \right\}$

$a'\underset{\sim}{x} - b \geq c$



If $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \Sigma$ are not known, but we have a "traing sample"

(n_1 observations $\underset{\sim}{x}_{1i}$, $i = 1, \ldots, n_1$ from π_1 and

n_2 observations $\underset{\sim}{x}_{2j}$, $j = 1, \ldots, n_2$ from π_2).

Then we estimate $\underset{\sim}{\mu}_1$ by $\overline{\underset{\sim}{x}}_1$

$\underset{\sim}{\mu}_2$ by $\overline{\underset{\sim}{x}}_2$.

Σ by Spooled, where

$$S_{pooled} = \frac{n_1-1}{n_1-1+n_2-1} S_1 + \frac{n_2-1}{n_1-1+n_2-1} S_2 \quad \text{and}$$

$$S_j = \frac{1}{n_j-1}\sum_{i=1}^{n_j}(\underset{\sim}{x}_{ji} - \overline{\underset{\sim}{x}}_j)(\underset{\sim}{x}_{ji} - \overline{\underset{\sim}{x}}_j)'.$$

We replace $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \Sigma$ by $\overline{\underset{\sim}{x}}_1, \overline{\underset{\sim}{x}}_2, S_{pooled}$ in R_1.

Note: $n_1-1+n_2-1 \geq p$.

Suppose π_1, π_2 are $N_p(\underset{\sim}{\mu}_1, \Sigma_1)$, $N_p(\underset{\sim}{\mu}_2, \Sigma_2)$, then

$$\frac{f_1(x)}{f_2(x)} = \frac{|\Sigma_2|^{1/2}}{|\Sigma_1|^{1/2}} \exp\left[-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma_1^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1) + \frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma_2^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)\right] = \frac{f_1(x)}{f_2(x)}$$

so, R_1

$$R_1 = \left\{ \underset{\sim}{x} : \frac{1}{2}\underset{\sim}{x}'(\Sigma_2^{-1}-\Sigma_1^{-1})\underset{\sim}{x} + (\underset{\sim}{\mu}_1'\Sigma_1^{-1} - \underset{\sim}{\mu}_2'\Sigma_2^{-1})\underset{\sim}{x} - \frac{1}{2}(\underset{\sim}{\mu}_1'\Sigma_1^{-1}\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2'\Sigma_2^{-1}\underset{\sim}{\mu}_2) + \frac{1}{2}\ln\frac{|\Sigma_2|}{|\Sigma_1|} \geq \ln\left[\frac{p_2\, c(1|2)}{p_1\, c(2|1)}\right] \right\}.$$

This is a quadratic separator.

When $\underset{\sim}{\mu}_1, \underset{\sim}{\mu}_2, \Sigma_1, \Sigma_2$ are not known, but we have ~~a~~ a training sample, then estimate $\underset{\sim}{\mu}_i$ by $\overset{\sim}{\bar{x}}_i$ and Σ_i by S_i.

Math B7800 5/8/18, Tue

Classification

Recall that in this set up the observations are coming from one of g possible groups: $\pi_1, \pi_2, \ldots, \pi_g$.

case 1: The pdf for different groups are known.

case 2: The pdfs' are unknown.

case 2.1: The pdfs' are parametric with unknown parameters.

case 2.2: The pdfs' are nonparametric.

In case 1, we minimize either ~~EPM~~ TMP or ECM (Total Misclassification probability) (Expected Cost of misclassification)

For $g=2$

$$\text{TMP} = p_1 \int_{R_2} f_1(x)\,dx + p_2 \int_{R_1} f_2(x)\,dx$$

p_i = prior probability of π_i.

$$\text{ECM} = p_1 c(2|1) \int_{R_2} f_1(x)\,dx + p_2 c(1|2) \int_{R_1} f_2(x)\,dx$$

where

Actual \ predict	π_1	π_2
π_1	0	$c(2\|1)$
π_2	$c(1\|2)$	0

The choice of R_1 and R_2 that minimizes ECM

$$R_1 = \left\{ \underset{\sim}{x} : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} > \frac{c(1|2)\, p_2}{c(2|1)\, p_1} \right\}$$

Note: if $c(2|1) = c(1|2)$, then minimizing ECM and TMP are equivalent.

$$R_2 = \left\{ \underset{\sim}{x} : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < \frac{c(1|2)\, p_2}{c(2|1)\, p_1} \right\}.$$

Optimum error rate (OER) = smallest possible error of a classification rule.

Ex Suppose π_1 and π_2 denote $N_p(\underset{\sim}{\mu}_1, \Sigma)$, $N_p(\underset{\sim}{\mu}_2, \Sigma)$ with known parameters. Suppose $c(2|1) = c(1|2)$, $p_1 = p_2 = \frac{1}{2}$.

Find OER?

The classification rule that minimizes ECM or TMP is given by

$$R_1 = \left\{ \underset{\sim}{x} : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq 1 \right\}, \quad R_2 = \left\{ \underset{\sim}{x} : \frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} < 1 \right\}$$

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} = \exp\left[-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1) + (\underset{\sim}{x}-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2) \right]$$

$$\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq 1 \iff \log\frac{f_1(\underset{\sim}{x})}{f_2(\underset{\sim}{x})} \geq 0 \iff (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} - (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\tfrac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) \geq 0$$

so, $R_1 = \left\{ \underset{\sim}{x} : (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} \geq (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\tfrac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) \right\}$

$R_2 = \left\{ \underset{\sim}{x} : (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} < (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\tfrac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) \right\}$.

"canonical dist"

$TMP = p_1 P(2|1) + p_2 P(1|2)$

$P(2|1) = \cancel{P(\underset{\sim}{x} \in R_2 \mid \underset{\sim}{x} \sim N(\underset{\sim}{\mu}_1, \Sigma))}$

$P(\underset{\sim}{x} \in R_2)$, where $\underset{\sim}{x} \sim N_p(\underset{\sim}{\mu}_1, \Sigma)$

so, $(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} \sim N_1\left((\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{\mu}_1, (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2) \right)$

$$P(\underset{\sim}{x} \in R_2) = P\left((\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} < (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\tfrac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) \right)$$

$$= P\left[\frac{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_1)}{\sqrt{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)}} < \frac{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\left[\frac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)-\underset{\sim}{\mu}_1\right]}{\sqrt{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)}} \right]$$

$$= P\left[Z < -\frac{1}{2}\underbrace{\sqrt{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)}}_{\Delta} \right]$$

$$= \Phi\left(-\frac{1}{2}\Delta\right)$$

$P(1|2) = P(\underset{\sim}{x} \in R_1)$ if $\underset{\sim}{x} \sim N_p(\underset{\sim}{\mu}_2, \Sigma)$.

$$= P\left[(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} \geq (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\tfrac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2) \right]$$, where

$(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{x} \sim N\left((\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\underset{\sim}{\mu}_2, (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2) \right)$

$$= P\left[\frac{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu}_2)}{\sqrt{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)}} \geq \frac{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}\left[\frac{1}{2}(\underset{\sim}{\mu}_1+\underset{\sim}{\mu}_2)-\underset{\sim}{\mu}_2\right]}{\sqrt{(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}(\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)}} \right] \quad (\underset{\sim}{\mu}_1-\underset{\sim}{\mu}_2)'\Sigma^{-1}$$

$= P\left[Z \geq \frac{1}{2}\Delta \right] = 1 - \Phi\left(\frac{1}{2}\Delta\right) = \Phi\left(-\frac{1}{2}\Delta\right)$. ← "OER"

$= P\left[Z \geq \frac{1}{2}\Delta\right] = 1 - \Phi\left(\frac{1}{2}\Delta\right) = \Phi\left(-\frac{1}{2}\Delta\right)$

So, $OER = \Phi\left(-\frac{\Delta}{2}\right)$

$\Delta = \sqrt{(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)' \underset{\sim}{\Sigma}^{-1} (\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)}$ is a "distance" between μ_1, μ_2.

This is called Mahalanobis (PC Mahalanobis) distance between two means.

Case 2.1 pdfs' are parametric with unknown parameters.

We have a training sample for which the group label is known.

Then we estimate the unknown parameters using the training sample.

Then use the classification rule used in case 1 with the parameters replaced with their estimates.

Here, we use $\hat{R}_1, \hat{R}_2$ instead of R_1 and R_2.

The corresponding rates are called Actual Error Rate (AER).

$$\text{AER} = p_1 \int_{\hat{R}_2} f_1(x)\,dx + p_2 \int_{\hat{R}_1} f_2(x)\,dx$$

AER is not known, as it involves the unknown pdfs' $f_1(x), f_2(x)$.

We estimate AER from the training data.

$\widehat{\text{AER}}$ (APER, apparent error rate)

$$= \frac{n_{1m} + n_{2m}}{n_1 + n_2} = \text{proportion of misclassification within training sample.}$$

"Training" Data	Actual \ Predicted	π_1	π_2	
	π_1	n_{1c}	n_{1m}	n_1
	π_2	n_{2m}	n_{2c}	n_2

Logistic Regression

Starting with training data for which group labels are known.

We can treat the group labels as the response variable.

For the ith training sample,

$$Y_i = \begin{cases} 1 & \text{if it is from } \pi_1 \\ 2 & \text{if it is from } \pi_2 \end{cases}$$

The data: $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ $(P \times 1)$

logistic model

$$\text{logit}\left(P(Y=1)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

$$= \log\left(\frac{P(Y=1)}{P(Y=0)}\right)$$

First estimate: $\beta_0, \beta_1, \ldots, \beta_p$.

Then allocate $\underset{\sim}{X}$ to π_1 if

$$\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p > 0.$$

~~Then allocate $\underset{\sim}{X}$ to π_1 if~~

~~$\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p > 0$.~~

When there are $g \geq 2$ populations $\pi_1, \ldots, \pi_g$ with pdf $f_1(x), \ldots, f_g(x)$.

$C(k|c)$ = cost of misclassification of an observation coming from π_i to π_k.

$p_1, p_2, \ldots, p_g$ are prior probabilities.

$$ECM(1) = C(2|1) f_2(x) + C(3|1) f_3(x) + \cdots + C(g|1) f_g(x)$$

$$\vdots$$

$$ECM(i) = \sum_{\substack{k=1 \\ k \neq i}}^{g} C(k|i) f_k(x)$$

$$ECM = p_1 ECM(1) + p_2 ECM(2) + \cdots + p_g ECM(g).$$

Allocate x to π_k if

$\sum_{\substack{i=1 \\ i \neq k}}^{g} p_i C(k|i) f_i(x)$ is the smallest.

Math B7800 — 5/10/18, Thur

· No date and time for variables on the project.

~~Last~~ Last time

- Classification with g groups / classes.
  - ↓ supervised learning

$\pi_1, \pi_2, \ldots, \pi_g$ are g groups with prior probabilities $p_1, p_2, \ldots, p_g$.

The pdfs are $f_1(x), \ldots, f_g(x)$

Case 1 : pdfs are known.

$ECM = p_1 ECM(1) + \cdots + p_g ECM(g)$.

cost of misclassification

$C(k|i)$ = cost for misclassifying an individual from π_i to π_k.

$$ECM(1) = C(2|1) \cdot P(2|1) + C(3|1) \cdot P(3|1) + \cdots + C(g|1) P(g|1)$$

$$\vdots$$

$$ECM(i) = \sum_{\substack{k=1 \\ k \neq i}}^{g} C(k|i) P(k|i), \quad i = 1, 2, \ldots, g.$$

$$ECM = \sum_{i=1}^{g} p_i \sum_{\substack{k=1 \\ k \neq i}}^{g} C(k|i) P(k|i)$$

$$= p_1 ECM(1) + \cdots + p_g ECM(g)$$

$P(k|i)$ = prob. of classificatying an observation from π_i to π_k

The classification rule that minimizes ECM will allocate x to π_k if $\sum_{\substack{i=1 \\ i \neq k}}^{g} f_i(x) p_i C(k|i)$ is minimum.

When $g=2$

we allocate x to π_1 if $f_1(x) P_1 C(2|1) \geq f_2(x) P_2 C(1|2)$ minimum.

$$\Leftrightarrow \frac{f_1(x)}{f_2(x)} \geq \frac{C(1|2) P_2}{C(2|1) P_1}$$

When the costs $C(k|i)$ are all equal, then it is enough to compare $\beta_k = \sum_{\substack{i=1 \\ i \neq k}}^{g} P_i f_i(x)$.

Choose k such that $\beta_k = \min_i \beta_i$

$$\beta_k = \sum_{i=1}^{g} P_i f_i(x) - P_k f_k(x)$$

$$= C - P_k f_k(x).$$

So, minimizing $\beta_k \Leftrightarrow$ maximizing $P_k f_k(x)$

Posterior prob. of a group

posterior prob. for group i

$= P(\pi_i | x) = P(X \text{ comes from } \pi_i \mid X = x)$

$$= \frac{P_i f_i(x)}{\sum_{k=1}^{g} P_k f_k(x)}$$

So, the allocation rule that maximizes the posterior prob. will allocate x to π_i if $P_i f_i(x) = \max_k P_k f_k(x)$

When the costs are equal, then minimizing ECM $\Leftrightarrow$ maximizing posterior.

Ex If π_i is $N_p(\underset{\sim}{\mu}_i, \Sigma_i)$, $i = 1, \ldots, g$.

Suppose equal costs. then Allocate $\underset{\sim}{x}$ to π_i if $P_i f_i(\underset{\sim}{x})$ is the max.

Equivalently, $\log(P_i f_i(\underset{\sim}{x}))$ is the max.

$$\Leftrightarrow \log(P_i) - \frac{1}{2}\log(|\Sigma_i|) - \frac{1}{2}(\underset{\sim}{x} - \underset{\sim}{\mu}_i)' \Sigma_i^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}_i) \quad \text{"quadratic"}$$

is the max.

Ex If π_i is $N_p(\underset{\sim}{\mu}_i, \Sigma)$, $i = 1, \ldots, g$, (Σ same)

then allocate $\underset{\sim}{x}$ to π_i if

$$\log(P_i) + \underset{\sim}{\mu}_i' \Sigma^{-1} \underset{\sim}{x} - \frac{1}{2} \underset{\sim}{\mu}_i' \Sigma^{-1} \underset{\sim}{\mu}_i \quad \text{"linear"}$$

is the max.

Chapter 12 Clustering (unsupervised learning).

No training data. ⟶ No fitted model. example 例子

Setup: we have a bunch of "objects", which we need to classify into different groups / clusters

Ex1 Community detection.

Ex2 Seggregating genes.

Ex3 Seggregating patients / healthy

First, assume that our observations are points in $\mathbb{R}^p$.

$\underset{\sim}{x}_1, \underset{\sim}{x}_2, \dots, \underset{\sim}{x}_n \in \mathbb{R}^p$ e.g. $p=2$

- Two kinds of algorithms

1. Heirarchical
2. Non-Heirarchical

1. Let $D_{n\times n}$ be the distance matrix, $D_{ij} = \|\underset{\sim}{x}_i - \underset{\sim}{x}_j\|$

Step 0 We have singleton clusters $\{\underset{\sim}{x}_1\}, \{\underset{\sim}{x}_2\}, \dots, \{\underset{\sim}{x}_n\}$

D = Distance matrix.

Step 1 find clusters U and V s.t. $\text{dist}(U,V)$ is the minimum among all distances of pairs of clusters.

Step 2 Merge clusters U and V. so delete U, V and add $U \cup V$.

[e.g. after step 0, if $\|\underset{\sim}{x}_1 - \underset{\sim}{x}_2\| = \min_{i\neq j} \|\underset{\sim}{x}_i - \underset{\sim}{x}_j\|$

then new cluster structure is $\{\underset{\sim}{x}_1, \underset{\sim}{x}_2\}, \{\underset{\sim}{x}_3\}, \dots, \{\underset{\sim}{x}_n\}$.

Step 3 Adjust the distance matrix.

Delete the rows and columns corresponding to clusters U and V.

Add a new row and column for $U \cup V$. in $D_{n-1\times n-1}$

Step 4 Repeat step 2 and 3 until you get one cluster $\{x_1, \dots, x_n\}$

Distance between clusters

(a) single link $\text{dist}(U,x) = \min\{d(x,y) : y \in U\}$.

(b) average link $\text{dist}(U,x) = \text{dist}(x, \text{centroid of } U)$

~~2. Non-~~

• Nonheirarchical Clustering

(k means clustering algorithm)

Assume that there are k clusters. $\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_k$

Goal: Find out the location of k centers and allocation of the points $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ to

$A(\underset{\sim}{x}_1), \ldots, A(\underset{\sim}{x}_n) \in \{\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_k\}$ so that

$\sum_{i=1}^{n} \| \underset{\sim}{x}_i - A(\underset{\sim}{x}_i) \|$ is minimum among all possible centers and all possible allocation rules. "Np hard"

so, "approximation"

Step1 Start with any k groups.

Step2 Obtain the k centroids these groups. (means)

Step3 Allocate each vector to its closest centroid. This defines new k groups.

Repeat Step 2 and 3 until the sum of the distances between the vectors and centroids stabilize.