19) p 230-3 Theory for the nonparametric prediction region for $\underline{y}_f$ given $\underline{x}_f$.

Suppose $\underline{y}_i = E\,\underline{y}_i + \underset{\sim}{\epsilon}_i = \hat{\underline{y}}_i + \hat{\underline{\epsilon}}_i$ where

$Cov(\underset{\sim}{\epsilon}_i) = cov(\underline{y}_i) = \Sigma_{\underset{\sim}{\epsilon}} > 0$ and the $\underset{\sim}{\epsilon}_i$ are iid.

Let $\hat{\underline{z}}_i = \hat{\underline{y}}_f + \hat{\underset{\sim}{\epsilon}}_i$ and $D_i^2(\hat{\underline{y}}_f, S_r) =$

$(\hat{\underline{z}}_i - \hat{\underline{y}}_f)^T S_r^{-1} (\hat{\underline{z}}_i - \hat{\underline{y}}_f)$.

$$\text{Let } q_n = \begin{cases} \min\left(1-\alpha+0.05,\ 1-\alpha+\frac{m}{n}\right) & \alpha > 0.1 \\ \min\left(1-\frac{\alpha}{2},\ 1-\alpha+\underbrace{10\alpha\frac{m}{n}}_{m \text{ not } p}\right), & \alpha \le 0.1 \end{cases}$$

If $q_n < 1-\alpha+.001$, set $q_n = 1-\alpha$.

Let $h = D_{(U_n)}$ where U_n is the q_nth sample quantile of the D_i. The $100(1-\alpha)\%$ nonparametric prediction region for $\underset{\sim}{y}_f$ is $\{\underline{z} \mid (\underline{z}-\hat{\underline{y}}_f)^T S_r^{-1}(\underline{z}-\hat{\underline{y}}_f) \le D^2_{(U_n)}\}$

$= \{\underline{z} \mid D_{\underline{z}}(\hat{\underline{y}}_f, S_r) \le D_{(U_n)}\}$.

a) Consider the n prediction regions for the data where $(\underline{y}_{fi}, \underline{x}_{fi}) = (\underline{y}_i, \underline{x}_i)$, $i = 1, \ldots, n$. If $D_{(U_n)}$ is unique, then "U_n" of the n prediction regions contain $\underline{y}_i$ where $\frac{U_n}{n} \to 1-\alpha$ as $n \to \infty$.

b) The nonparametric prediction region is a large sample $100(1-\alpha)\%$ prediction region for $\underline{y}_f$ if $(\hat{\underline{y}}_f, S_r)$ is a consistent estimator of $(E(\underline{y}_f), \Sigma_{\underset{\sim}{\epsilon}})$.

c) If b) holds and the iid $\underset{\sim}{\epsilon}_i$ come from a large class of elliptically contoured distributions such that the highest density region is $\{\underline{z} \mid D_{\underline{z}}(\underline{0}, \Sigma_{\underset{\sim}{\epsilon}}) \le D_{1-\alpha}\}$ then the nonparametric region is asymptotically optimal (has smallest asymptotic volume with coverage $\to 1-\alpha$).

proof a) Let $(\underline{x}_f, \underline{y}_f) = (\underline{x}_i, \underline{y}_i)$. Let $\underline{S_r = \hat{\Sigma}_{\underset{\sim}{\epsilon}}}$.

$$D^2_{\underline{y}_i}(\hat{\underline{y}}_f, \hat{\Sigma}_{\underset{\sim}{\epsilon}}) = (\underline{y}_i - \hat{\underline{y}}_i)^T \hat{\Sigma}_{\underset{\sim}{\epsilon}}^{-1} (\underline{y}_i - \hat{\underline{y}}_i) = \hat{\underline{\epsilon}}_i^T \hat{\Sigma}_{\underset{\sim}{\epsilon}}^{-1} \hat{\underline{\epsilon}}_i = D^2_{\hat{\underset{\sim}{\epsilon}}_i}(\underline{0}, \hat{\Sigma}_{\underset{\sim}{\epsilon}}).$$

Hence $\underline{y}_i$ is in the ith prediction region $\{\underline{z} \mid D_{\underline{z}}(\hat{\underline{y}}_i, \hat{\Sigma}_{\underset{\sim}{\epsilon}}) \le D_{(U_n)}(\hat{\underline{y}}_i, \hat{\Sigma}_{\underset{\sim}{\epsilon}})\}$ iff $\hat{\underline{\epsilon}}_i$ is in the prediction region $\{\underline{z} \mid D_{\underline{z}}(\underline{0}, \hat{\Sigma}_{\underset{\sim}{\epsilon}}) \le D_{(U_n)}(\underline{0}, \hat{\Sigma}_{\underset{\sim}{\epsilon}})\}$, but "$U_n$" of the $\hat{\underline{\epsilon}}_i$ are in this region if $D_{(U_n)}$ is unique.

MU 5.2

Since $D_{(U_n)}$ is the $1-\alpha$ percentile of the D_i asymptotically, $\frac{U_n}{n} \to 1-\alpha$.

b) Let $P[D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})] = 1-\alpha$.

Since $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}} > 0$, prop 5.1 shows that if

$(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} (E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, then

$D(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \xrightarrow{P} D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$. Hence the percentiles of the distances converge in probability, and $P(\boldsymbol{y}_f \in \{\boldsymbol{z} \mid D_{\boldsymbol{z}}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}}) \le D_{1-\alpha}(\hat{\boldsymbol{y}}_f, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}})\})$

$\to P[\boldsymbol{y}_f \in \{\boldsymbol{z} \mid D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}] = 1-\alpha$.

c) The asymptotically optimal prediction region with smallest volume (hence highest density) such that coverage is $1-\alpha$ as $n \to \infty$ is $\{\boldsymbol{z} \mid D_{\boldsymbol{z}}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\alpha}(E(\boldsymbol{y}_f), \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$ if the asymptotically optimal region for the $\boldsymbol{\epsilon}_i$ is

$\{\boldsymbol{z} \mid D_{\boldsymbol{z}}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}) \le D_{1-\alpha}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\}$. So the result follows by b).

20) SUR is Not on exams \$12.7 econometrics 52.5

20) p235 The seemingly unrelated regressions (SUR) model is like the multivariate linear regression model except

$$\underset{\sim}{Y_j} = \mathbb{X}_j \underset{\sim}{\beta_j} + \underset{\sim}{e_j} \qquad j = 1, \ldots, m.$$

So there is a different $\mathbb{X}_j$ for each response variable Y_j.

Let $\underset{\sim}{\beta_j}$ be $k_j \times 1$ and $\underset{\sim}{x_{ij}} = (1, x_{2j}, \ldots, x_{k_j, j})^T$.

Then the ith case is

$$(Y_{i1}, \ldots, Y_{im}, x_{21}, \ldots x_{k_1,1}, x_{22}, \ldots, x_{k_2,2}, x_{2m}, \ldots, x_{k_m,m})^T,$$

omitting the m ones.

21) The multivariate linear regression model is a special case of SUR if

$$\mathbb{X}_j \equiv \mathbb{X} \text{ so } \underset{\sim}{x_{ij}} \equiv \underset{\sim}{x_i} \qquad j = 1, \ldots, m.$$

22) The SUR model is a special case of the mreg model that has $\underset{\sim}{x_i} = (1, x_{21}, \ldots, x_{k_m,m})^T$.

Let $k = \sum_{i=1}^{m} k_i$.

Then X is $n \times (\underbrace{k - m + 1}_{\text{use one 1, not } m})$

and $\underset{\sim}{\beta}_j^* = \begin{pmatrix} \beta_{1j} \\ 0 \\ \vdots \\ 0 \\ \beta_{2j} \\ \vdots \\ \beta_{kj} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ is the jth column of B with most entries $= 0$. The nonzero entries form $\underset{\sim}{\beta}_j$ from the SUR model.

23) p236 The SUR model is

$$\underset{\sim}{y}_i = E\,\underset{\sim}{y}_i + \underset{\sim}{\varepsilon}_i = \begin{pmatrix} \underset{\sim}{x}_{i1}^T \underset{\sim}{\beta}_1 \\ \underset{\sim}{x}_{i2}^T \underset{\sim}{\beta}_2 \\ \vdots \\ \underset{\sim}{x}_{im}^T \underset{\sim}{\beta}_m \end{pmatrix} + \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{im} \end{pmatrix}$$

$$= \begin{pmatrix} \underset{\sim}{x}_{i1}^T \hat{\underset{\sim}{\beta}}_1 \\ \underset{\sim}{x}_{i2}^T \hat{\underset{\sim}{\beta}}_2 \\ \vdots \\ \underset{\sim}{x}_{im}^T \hat{\underset{\sim}{\beta}}_m \end{pmatrix} + \begin{pmatrix} \hat{\varepsilon}_{i1} \\ \vdots \\ \hat{\varepsilon}_{im} \end{pmatrix} = \hat{\underset{\sim}{y}}_i + \hat{\underset{\sim}{\varepsilon}}_i \text{ and } \mathrm{Cov}(\varepsilon_i) = \Sigma_{\underset{\sim}{\varepsilon}}.$$

24) The SUR estimators $\hat{\underset{\sim}{\beta}}_j$, $j = 1, \ldots, m$, and $\hat{\Sigma}_{\underset{\sim}{\varepsilon}}$ are thought to be better than the least squares estimators.

25] Make the m response and residual plots of $\hat{Y}_j$ vs Y_j and $\hat{Y}_j$ vs r_j $j = 1, ..., m$ models $i = 1, ..., n$

and the DD plot of the $\hat{\underset{\sim}{\varepsilon}}_i$.

53.5

26] SUR prediction region for $\underline{y}_f$ given $\underline{x}_{f,1}, ..., \underline{x}_{f,m}$

let $\hat{\underline{z}}_i = \hat{\underline{y}}_f + \hat{\underset{\sim}{\varepsilon}}_i \quad i = 1, ..., n$

where $\hat{\underline{y}}_f = (\underline{x}_{f,1}^T \hat{\underline{\beta}}_1, ..., \underline{x}_{f,m}^T \hat{\underline{\beta}}_m)^T$.

The prediction region is

$$\{\underline{z} \mid D_{\underline{z}}(\hat{\underline{y}}_f, \hat{\underline{\Sigma}}_{\underset{\sim}{\varepsilon}}) \le D_{(U_P)}(\hat{\underline{y}}_f, \hat{\underline{\Sigma}}_{\underset{\sim}{\varepsilon}})\}$$

where U_P is as before.

Note that $(\underline{y}_i - \hat{\underline{y}}_f)^T \hat{\underline{\Sigma}}_{\underset{\sim}{\varepsilon}}^{-1} (\underline{y}_i - \hat{\underline{y}}_f) =$

$\hat{\underset{\sim}{\varepsilon}}_i^T \hat{\underline{\Sigma}}_{\underset{\sim}{\varepsilon}}^{-1} \hat{\underset{\sim}{\varepsilon}}_i$ so $\underline{y}_i$ is in its

prediction region (with $\underline{x}_{f,j} = \underline{x}_{ij}$ $j = 1, ..., m$)

if $\hat{\underset{\sim}{\varepsilon}}_i$ is in the region $\{\underline{z} \mid D_{\underline{z}}(\underline{0}, \hat{\underline{\Sigma}}_{\underset{\sim}{\varepsilon}}) \le D_{(U_P)}(\underline{0}, \hat{\underline{\Sigma}}_{\underset{\sim}{\varepsilon}})\}$.

But this is not the nonparametric region since the sample mean of the SUR $\hat{\underset{\sim}{\varepsilon}}_i$ is not $\underline{0}$.

Ch13 1) Factor analysis grouped highly correlated variables X_j together (columns of the data matrix).

Clustering groups cases $\underline{x}_i$ together (rows of the data matrix).

2) Discriminant analysis is called supervised classification while clustering is called unsupervised classification: there are no labelled responses (groups for discriminant analysis).

3) K means clustering:

i) partition the n cases into k initial groups and find the mean of each group. Alternatively choose k initial seed points. These are groups of size 1 so the mean is equal to the seed point.

ii) compute distances between each case and each mean. Assign a case to the cluster whose mean is nearest.

iii) recalculate the mean of each cluster. (S4.5

iv) go to ii) and repeat until no changes occur.

4) 2 problems i) there could be more or less than K clusters

ii) 2 initial seedpoints could actually belong to the same cluster.

5) Hierarchical Clustering needs a distance.

Single linkage is the minimum distance between cases in cluster i and cases in cluster j. - nearest neighbor

Complete linkage is the maximum distance

single linkage is the minimum distance.

Sometimes the average distance is used.

i) Start with $m=n$ clusters so each case forms a cluster. Compute the distance matrix for the n clusters. Let $d_{u,v}$ be the smallest distance. Combine clusters U and V into a single cluster and set $m=n-1$.

ii) repeat i) with the new m. Continue until there is a single cluster.

iii) Plot the results in a ~~dendogram~~ dendrogram. Use the