16) The prediction region method large sample $100(1-\delta)\%$ confidence region for $\underline{\theta}$ is

$$\{\underline{w}: \underbrace{(\underline{w}-\underline{T}^*)^T [S_T^*]^{-1} (\underline{w}-\underline{T}^*)}_{D^2_{\underline{w}}(\underline{T}^*, S_T^*)} \leq D^2_{(U_B)}\}$$

where $D_i^2 = (\underline{T}_i^* - \underline{T}^*)^T [S_T^*]^{-1} (\underline{T}_i^* - \underline{T}^*)$.

Reject $H_0: \underline{\theta} = \underline{\theta}_0$ if $(\underline{T}^* - \underline{\theta}_0)^T [S_T^*]^{-1} (\underline{T}^* - \underline{\theta}_0) > D^2_{(U_B)}$.

17) The (modified) Bickel and Ren large sample $100(1-\delta)\%$ confidence region is

$$\{\underline{w}: (\underline{w}-\underline{T}_n)^T (S_T^*)^{-1} (\underline{w}-\underline{T}_n) \leq D^2_{(U_B, T)}\}$$

where U_B is the $100(1-\delta)$th percentile

where $D_i^2 = (\underline{T}_i^* - \underline{T}_n)^T (S_T^*)^{-1} (\underline{T}_i^* - \underline{T}_n)$.

18) The hybrid large sample $100(1-\delta)\%$ confidence region is

$$\{\underline{w}: (\underline{w}-\underline{T}_n)^T (S_T^*)^{-1} (\underline{w}-\underline{T}_n) \leq D^2_{(U_B)}\}.$$

19) 16) is the nonparametric prediction region applied to the bootstrap sample.

20) Ch 2 applied 2 PIs to the bootstrap sample to get large sample CIs.

21) The confidence regions tend to work (25.5)
if $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$
where $E(U) = 0$ and $Cov(U) = \Sigma_U > 0$.
(positive definite)

WRT the bootstrap distr, T_n is a constant,

and $\sqrt{n}\begin{pmatrix} T_1^* - T_n \\ \vdots \\ T_B^* - T_n \end{pmatrix} \xrightarrow{D} \begin{pmatrix} V_1 \\ \vdots \\ V_B \end{pmatrix}$, $V_i \overset{iid}{\sim} U$. Multiply

both sides by $A = \left(\frac{1}{B} \cdots \frac{1}{B}\right)$ to take the sample mean of both sides.

For fixed B, get $\sqrt{n}(\bar{T}^* - T_n) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} V_i$

As $B \to \infty$, $\sqrt{n}(\bar{T}^* - T_n) \xrightarrow{D} 0$ so $\bar{T}^*$ and T_n are asymptotically equivalent and $\sqrt{n}(\bar{T}^* - \theta)$ —

$= \sqrt{n}(\bar{T}^* - T_n + T_n - \theta) = \sqrt{n}(\bar{T}^* - T_n) + \sqrt{n}(T_n - \theta) \xrightarrow{D} 0 + U = U$

22) comparing the iid data cloud $T_1, \ldots, T_B$ with the bootstrap data cloud $T_1^*, \ldots, T_B^*$ is useful where $T_i = T_{in}$. If $\sqrt{n}(T_n - \theta) \xrightarrow{D} U$ and $\sqrt{n}(T_i^* - T_n) \xrightarrow{D} U$, then the bootstrap data cloud is like the iid data cloud (which is centered at θ) shifted to be centered at T_n.

		Zemen, Samuel G.

23} Geometric argument: suppose

$\sqrt{n}(T_n - \underline{\theta}) \xrightarrow{D} \underline{U}$ with $E(\underline{U}) = \underline{0}$ and $Cov(\underline{U}) = \Sigma_U \neq 0$.

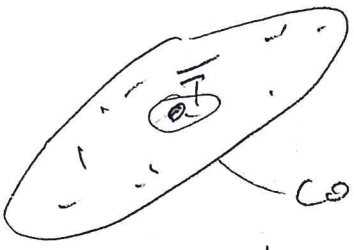
Assume $T_1, \ldots, T_B$ are iid with nonsingular covariance matrix Σ_{T_A}. Assume $(n S_T)^{-1} \xrightarrow{P} \Sigma_A^{-1}$.

The nonparametric prediction region applied to $T_1, \ldots, T_B$ is $R_p = \{\underline{w}: D^2_{\underline{w}}(\overline{T}, S_T) \leq D^2_{(U_B)}\}$ is centered at $\overline{T}$ and contains a future value of the statistic T_E with prob $1 - \delta_B \to 1 - \delta$ as $B \to \infty$ (eventually bounded below).

The region $R_C = \{\underline{w}: D^2_{\underline{w}}(T_n, S_T) \leq D^2_{(U_B)}\}$ is a large sample confidence region for $\underline{\theta}$ where T_n is a randomly selected T_i.

Proof} R_C contains $\overline{T}$ with prob $1 - \delta_B$

$D_{T_i}(\overline{T}, S_T) = D_{\overline{T}}(T_i, S_T)$ ($\overline{T}$ centered at $\overline{T}$; T_i centered at T_i)



contains $100(1-\delta_B)\%$ of the T_i

Any T_i in the hyperellipsoid R_p has R_C that contains $\overline{T}$. Any T_i outside of R_p has R_C that does not contain $\overline{T}$.

Ex B, $\sqrt{n}(\overline{T} - \theta) \xrightarrow{D} \frac{1}{B}\sum_{i=1}^{B} \underline{U}_i$, $\underline{U}_i \overset{iid}{\sim} \underline{U}$. So $\overline{T}$ gets arbitrarily close to $\underline{\theta}$ as $B \to \infty$ compared to the T_i.

So R_C contains $\underline{\theta}$ with prob $\to 1 - \delta_B \to 1 - \delta$.

contains $\underline{\theta}$ with 99% prob is not a point. (24)

So the coverage tends to be less than $1-\delta$ if $g_B = \sqrt{B}$ is [illegible]. Increasing the coverage using g_B makes the observed coverage closer to the nominal $1-\delta$.

25) The bootstrap data cloud is like the iid data cloud shifted to be centered at T_n. So applying the nonparametric prediction region to the bootstrap data cloud gives a confidence region.

26) Suppose $\sqrt{n}(T_n - \underline{\theta}) \xrightarrow{D} \underline{0}$, $\sqrt{n}(\overline{T}_i^* - T_n) \xrightarrow{D} \underline{0}$, $\sqrt{n}(\overline{T}_i^* - \overline{T}^*) \xrightarrow{D} \underline{0}$ and $\sqrt{n}(\overline{T}^* - \underline{\theta}) \xrightarrow{D} \underline{0}$. see 23). Then

$$D_1^2 = D_{\overline{T}_i^*}^2(\overline{T}^*, S_T^*) = \sqrt{n}(\overline{T}_i^* - \overline{T}^*)(nS_T^*)^{-1}\sqrt{n}(\overline{T}_i^* - \overline{T}^*)$$

$$D_2^2 = D_{\underline{\theta}}^2(T_n, S_T^*) = \sqrt{n}(T_n - \underline{\theta})^T(nS_T^*)^{-1}\sqrt{n}(T_n - \underline{\theta})$$

$$D_3^2 = D_{\underline{\theta}}^2(\overline{T}^*, S_T^*) = \sqrt{n}(\overline{T}^* - \underline{\theta})(nS_T^*)^{-1}\sqrt{n}(\overline{T}^* - \underline{\theta})$$

$$D_4^2 = D_{\overline{T}_i^*}^2(T_n, S_T^*) = \sqrt{n}(\overline{T}_i^* - \overline{T}^*)(nS_T^*)^{-1}\sqrt{n}(\overline{T}_i^* - \overline{T}^*)$$

If $(nS_T^*)^{-1} \xrightarrow{P} \Sigma_A^{-1}$, then $D_j^2 \xrightarrow{D} \underline{U}^T \Sigma_A^{-1} \underline{U}$. If $(nS_T^*)^{-1}$ is "not too ill conditioned," then $D_j^2 \approx \underline{U}^T(nS_T^*)^{-1}\underline{U}$ for large n. Then the 3 bootstrap confidence regions will have coverage near $1-\delta$.

27} Data splitting divides the training data $\underline{x}_1, \ldots, \underline{x}_n$ into two sets H and V where H has $n_H \geq \frac{n}{2}$ of the cases and V has $n - n_H = n_V$ of the cases. The estimator (T_H, C_H) is computed using the data set H. Then the squared validation distances

$$D_j^2 = D^2_{\underline{x}_j}(T_H, C_H) = (\underline{x}_j - T_H)^T C_H^{-1} (\underline{x}_j - T_H)$$

are computed for the $j = 1, \ldots, n_V$ cases in the validation set V. Let $D^2_{(U_V)}$ be the U_Vth order statistic of the D_j^2 where

$$U_V = \min(n_V, \lceil (n_V + 1)(1-\delta) \rceil).$$

28} The large sample $100(1-\delta)\%$ data splitting prediction region for $\underline{x}_f$ is

$$\{\underline{z} : D^2_{\underline{z}}(T_H, C_H) \leq D^2_{(U_V)}\}.$$

29} Can use $(T_H, C_H) = (T_{RMVN}, C_{RMVN})$, or, if $p > n$, $(T_H, C_H) = (\underline{\bar{x}}, I)$ or $(MED(W), I)$.

30} $D_{(1)}, \ldots, D_{(n)}$ have rank $1, \ldots, n$.

Consider D_k^2 for $k=1,\dots,n_v, n_v+1$ where $D_{n_v+1}^2$ is for $\underline{x}_t$. Since the n_v+1 cases are iid, the prob that D_t^2 has rank j for $j=1,\dots,n_v+1$ is $\frac{1}{n_v+1}$ for each t. So the ranks follow the discrete uniform distribution. (39.5)

Let $D_{(j)}^2$ be the order statistics using $j=1,\dots,n_v$ without using the unknown $D_{n_v+1}^2$. Then $D_{(i)}^2$ has rank i if $D_{(i)}^2 < D_{n_v+1}^2$ but rank $i+1$ if $D_{(i)}^2 > D_{n_v+1}^2$. Thus $D_{(v)}^2$ has rank $v+1$ if $D_{\underline{x}_t}^2 < D_{(v)}^2$ and

$$P\left(\underline{x}_t \in \{\underline{z} : D_{\underline{z}}^2(T_H, C_H) \le D_{(v)}^2\}\right) = P\left(D_{\underline{x}_t}^2 \le D_{(v)}^2\right)$$

($P(\underline{x}_t \text{ has rank} \le v) = P(\underline{x}_t \text{ has rank} < v+1)$)

$$\ge \frac{v}{n_v+1} \to 1-\delta \text{ as } n_v \to \infty.$$

$=$ if there are no ties.

32) =.05

n_v	1	2	19	20	39	59	79	99
$\frac{v}{n_v+1}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{19}{20}=.95$	$\frac{20}{21}$	$\frac{38}{40}=.95=\frac{57}{60}$	$\frac{76}{80}=.95=\frac{95}{100}$		
	↑ so so chance							

33) Need $n_v \ge 100$ to estimate the $100(1-\delta)$ percentile of the D_j^2, $.05 \le \delta \le .5$. Volume can be huge for small n_v.

or if (T_A, C_A) is bad.

1] The multiple linear regression (MLR) model

$$Y_i = x_{i,1}\beta_1 + \cdots + x_{ip}\beta_p + e_i = \underline{x}_i^T \underline{\beta} + e_i \quad i = 1, \ldots, n$$

In matrix form $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underline{e}$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \quad \text{where}$$

$$\underset{\sim}{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \\ x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

2] MLR is important for Math 484, 584 (linear model theory) 586 (Statistical learning) and robust statistics.

3] a) The constant variance MLR model has the e_i iid, $E(e_i) = 0$, $V(e_i) = \sigma^2$.

b) The unimodal MLR model is as in a) but adds the assumption that the e_i are from a unimodal distribution that is not highly skewed.

c) The normal MLR model has $e_i \overset{iid}{\sim} N(0, \sigma^2)$.

for now skim § 5.1, 5.2 / 2.5.

§ 5.4

4} To get an asymptotically optimal PI for Y_f given $\underline{x}_f$ and $(Y_1, \underline{x}_1), \ldots, (Y_n, \underline{x}_n)$,

$\underline{x}_f$: test data; $(Y_1, \underline{x}_1), \ldots, (Y_n, \underline{x}_n)$: training data

find the shorth(c) estimator $[\hat{r}_{(s)}, \hat{r}_{(s+c-1)}]$ of the residuals and use

$$[\hat{Y}_f + a_n \hat{r}_{(s)}, \hat{Y}_f + a_n \hat{r}_{(s+c-1)}].$$

$\hat{Y}_f \xrightarrow{p} E(Y_f | \underline{x}_f)$ and $[\hat{r}_{(s)}, \hat{r}_{(s+c-1)}]$ estimates the pop shorth of the iid e_i.

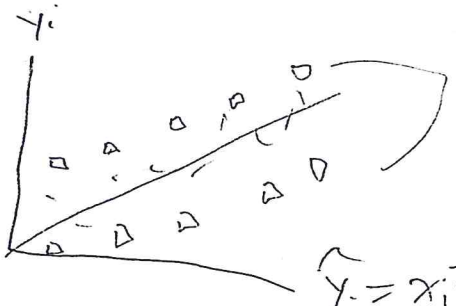
Here $a_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n}{n-p}}\sqrt{1+h_f}$ where

$r_i \approx \sqrt{\frac{n}{n-p}}\, e_i$ in that $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^n r_i^2 \approx \frac{1}{n}\sum_{i=1}^n e_i^2$

and $h_f = \underline{x}_f^T (X^T X)^{-1} \underline{x}_f$ is the leverage of $\underline{x}_f$.

Want $h_f \le \max \underline{x}_i^T (X^T X)^{-1} \underline{x}_i = h_i$ where

h_i is the ith diagonal element of $H = X(X^T X)^{-1} X^T$.



response plot

PI limits not quite parallel to identity line

Skim §1.5