12) Low dimensions for some regression model. $\underline{\beta} = \underline{\beta}(\underline{x}, Y) = \underline{\beta}_F$ = full model.

Assume $\sqrt{n}(\hat{\underline{\beta}} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, V)$.

Typically if $S \subseteq I$, then

$$\sqrt{n}\,(\underbrace{\hat{\underline{\beta}}_I - \underline{\beta}_I}_{a \times 1}) \xrightarrow{D} N_a(\underline{0}, V_I)$$

where usually V_I can <u>not</u> be obtained by taking the appropriate elements of V ($\hat{\underline{\eta}} = \hat{\underline{\Sigma}}_{xY}$ was an exception).

Then $\sqrt{n}(\hat{\underline{\beta}}_{I,0} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, V_{I,0})$

where V_{I0} adds rows and columns of 0's corresponding to the omitted variables. Hence V_{I0} is singular unless $I = F = \{1, \ldots, p\}$ corresponds to the full model $\underline{\beta}_F = \underline{\beta}$.

end exam 1 material
~~exam 1 material~~

13] If $S \not\subseteq I$, could have

$$\sqrt{n}\left(\hat{\underset{\sim}{\beta}}_I - \underset{\sim}{\beta}_I\right) \xrightarrow{D} N_a\left(\underset{\sim}{0}, V_I\right)$$ this could fail, too

(multitude of models)

but $\hat{\underset{\sim}{\beta}}_{I,0}$ is no longer a consistent estimator of $\underset{\sim}{\beta} = \underset{\sim}{\beta}_F$. So $\sqrt{n}(\hat{\underset{\sim}{\beta}}_{I,0} - \underset{\sim}{\beta})$ does not converge in distribution. ($\underset{\sim}{\beta}_{I,0} \neq \underset{\sim}{\beta} = \underset{\sim}{\beta}_F$)

ex] $p = 4$, $S = \{1\}$, $I = \{1, 3\}$

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_3\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_3\end{pmatrix}\right] \xrightarrow{D} N_2\left[\underbrace{\begin{pmatrix}0\\0\end{pmatrix}}_{\underset{\sim}{0}}, \begin{pmatrix}V_{11} & V_{13}\\ V_{31} & V_{33}\end{pmatrix}\right]$$

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_1\\ 0\\ \hat{\beta}_3\\ 0\end{pmatrix} - \underbrace{\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}}_{\underset{\sim}{\beta}}\right] \xrightarrow{D} N_4\left[\underbrace{\begin{pmatrix}0\\0\\0\\0\end{pmatrix}}_{\underset{\sim}{0}}, \underbrace{\begin{pmatrix}V_{11} & 0 & V_{13} & 0\\ 0 & 0 & 0 & 0\\ V_{31} & 0 & V_{33} & 0\\ 0 & 0 & 0 & 0\end{pmatrix}}\right]$$

$$\underset{\sim}{\beta} = \underset{\sim}{\beta}_F = \begin{pmatrix}\beta_1\\ 0\\ 0\\ 0\end{pmatrix}$$ so $\beta_3 = 0$.

Rows and columns corresponding to 2 and 3 are 0's.

14) A random vector $\underline{u}$ has a mixture distribution of random vectors $\underset{\sim}{u}_j$ with probs π_j if $\underline{u}$ equals randomly selected vector $\underline{u}_j$ with prob's π_j for $j=1,\ldots,J$. (The selection process must not change the dist of $\underline{u}_j$.) Then the cdf

$$F_{\underset{\sim}{u}}(\underset{\sim}{t}) = \sum_{j=1}^{J} \pi_j F_{\underset{\sim}{u}_j}(\underset{\sim}{t})$$

where

$0 \le \pi_j \le 1$, $\sum_{j=1}^{J} \pi_j = 1$, $J \ge 2$, and

$F_{\underline{u}_j}(\underset{\sim}{t})$ is the cdf of $\underset{\sim}{u}_j$.

Suppose $E(h(\underline{u}))$ and $E[h(\underline{u}_j)]$ exist.

Then $E[h(\underline{u})] = \sum_{j=1}^{J} \pi_j E[h(\underline{u}_j)]$,

$$E[\underline{u}] = \sum_{j=1}^{J} \pi_j E(\underline{u}_j), \quad \text{and}$$

$$\operatorname{cov}(\underline{u}) = \sum_{j=1}^{J} \pi_j \operatorname{cov}(\underline{u}_j) + \sum_{j=1}^{J} \pi_j E(\underline{u}_j)[E(\underset{\sim}{u}_j)]^T - E(\underline{u})[E(\underline{u})]^T.$$

$E[\underline{u}_j \underline{u}_j^T] = \operatorname{cov}(\underline{u}_j) + E(\underline{u}_j)[E(\underline{u}_j)]^T$

$\operatorname{cov}(\underline{u}) = E[\underline{u}\,\underline{u}^T] - E(\underline{u})[E(\underline{u})]^T$

If $E(\underset{\sim}{u}_j) = \underset{\sim}{0}$ for $j = 1, \ldots, J$, then

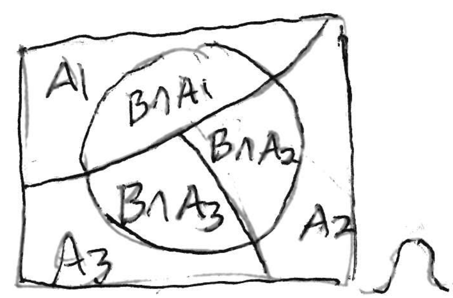
$$E(\underset{\sim}{u}) = \underset{\sim}{0} \text{ and } Cov(\underset{\sim}{u}) = \sum_{j=1}^{J} \pi_j \, Cov(\underset{\sim}{u}_j).$$

15] Law of Total Probability:

Let $A_1, \ldots, A_J$ form a partition of the sample space Ω (the A_i are disjoint, $P(A_i) > 0$, and $\bigcup_{i=1}^{J} A_i = \Omega$). Then

$$P(B) = \sum_{K=1}^{J} P(B \cap A_K) = \sum_{K=1}^{J} P(B \mid A_K)\, P(A_K).$$

variant: If $P(A_K) = 0 = P(B \cap A_K)$, define $P(B \mid A_K)\, P(A_K) = 0$.



16] Let $P(\hat{\underset{\sim}{\beta}}_{VS} = \hat{\underset{\sim}{\beta}}_{I_{K},0}) = \pi_{kn}$ for $k = 1, \ldots, J$.

Let $\underset{\sim}{w}_n = \sqrt{n}(\hat{\underset{\sim}{\beta}}_{VS} - \underset{\sim}{\beta})$.

Let cdf $F_{\underset{\sim}{z}}(\underset{\sim}{t}) = P(z_1 \le t_1, \ldots, z_p \le t_p)$.

Let $\hat{\underline{\beta}}^{c}_{Ik,0}$ be a random vector from the conditional dist $\hat{\underline{\beta}}_{Ik,0} \mid (\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{Ik,0})$.

Let $\underline{w}_{kn} = \sqrt{n}(\hat{\underline{\beta}}^{c}_{Ik,0} - \underline{\beta}) \overset{D}{=} \sqrt{n}(\hat{\underline{\beta}}_{Ik,0} - \underline{\beta}) \mid (\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{Ik,0})$.

Then $F_{\underline{w}_n}(\underline{t}) = \sum_{k=1}^{J} F_{\underline{w}_{kn}}(\underline{t})\, \pi_{kn}$.

Hence $\hat{\underline{\beta}}_{VS}$ has a mixture dist of the $\hat{\underline{\beta}}_{Ik,0}$ with prob's π_{kn} and $\underline{w}_n$ has a mixture dist of the $\underline{w}_{kn}$ with prob's π_{kn}.

Law of Total Prob (variant)

proof) $F_{\underline{w}_n}(\underline{t}) = P[\underbrace{\sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta}) \le \underline{t}}_{B}] =$

$$\sum_{k=1}^{J} P(\sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta}) \le \underline{t} \mid \underbrace{\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{Ik,0}}_{A_k})\, \underbrace{P(\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{Ik,0})}_{\pi_{kn}}$$

$$= \sum_{k=1}^{J} P\left(\sqrt{n}\left(\hat{\underline{\beta}}_{I_k,0} - \underline{\beta}\right) \le \underline{t} \mid \hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_k,0}\right) \pi_{kn}$$

$$= \sum_{k=1}^{J} P\left(\sqrt{n}\left(\hat{\underline{\beta}}^{C}_{I_k,0} - \underline{\beta}\right) \le \underline{t}\right) \pi_{kn}$$

$$= \sum_{k=1}^{J} F_{\underline{w}_{kn}}(\underline{t})\, \pi_{kn}. \quad \square$$

16] Let $\hat{\underline{\beta}}_{MIX} = \hat{\underline{\beta}}_{I_k,0}$ with prob π_{kn}, but let the selection process be ind of $\hat{\underline{\beta}}_{I_k,0}$ so that $P[\sqrt{n}(\hat{\underline{\beta}}_{MIX} - \underline{\beta}) \le \underline{t}) =$

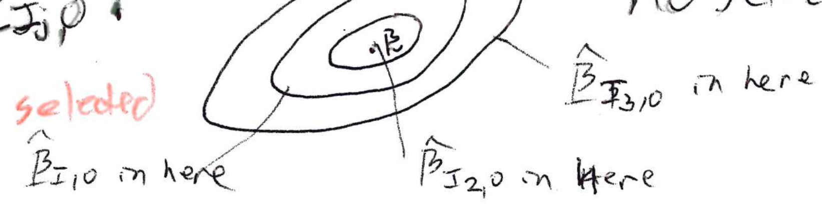
no C because of ind

$$\sum_{k=1}^{J} P\left(\sqrt{n}(\hat{\underline{\beta}}_{I_k,0} - \underline{\beta}) \le \underline{t} \mid \hat{\underline{\beta}}_{MIX} = \hat{\underline{\beta}}_{I_k0}\right)\pi_{kn} = \sum_{k=1}^{J} P\left[\sqrt{n}(\hat{\underline{\beta}}_{I_k0} - \underline{\beta}) \le \underline{t}\right] \pi_{kn}$$

can't compute $\hat{\underline{\beta}}_{MIX}$ from data, but to simulate $\hat{\underline{\beta}}_{MIX}$, generate full model data set and perform variable selection. If $\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_j,0}$, generate another full model data set, and fit $\underbrace{\hat{\underline{\beta}}_{I_j}}_{\text{no selection bias}}$.

Then $\hat{\underline{\beta}}_{MIX} = \hat{\underline{\beta}}_{I_j,0}$.

VS makes vectors far from $\underline{\beta}$ unlikely to be selected



17] Let $\underline{u}_n = \sqrt{n}\,(\hat{\underline{\beta}}_{MIX} - \underline{\beta})$

and let $\underline{u}_{kn} = \sqrt{n}\,(\hat{\underline{\beta}}_{I_{k},0} - \underline{\beta})$.

Then $F_{\underline{u}_n}(\underline{t}) = \sum_{k=1}^{J} F_{\underline{u}_{kn}}(\underline{t})\,\pi_{kn}$ by

nearly the same proof as 15].

18] If $S \subseteq I_j$ where $\dim(I_j) = a_j$,

then $\sqrt{n}\,(\hat{\underline{\beta}}_{I_j} - \underline{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\underline{0}, V_j)$ and

$\underline{u}_{jn} = \sqrt{n}\,(\hat{\underline{\beta}}_{I_j,0} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, V_{j0}) = \underset{\sim}{u}_j$

where V_{j0} adds rows and columns of 0's corresponding to the x_k not in I_j.

see 12].

19] If all of the $F_{\underline{u}_{kn}}(\underline{t}) \to F_{\underline{u}_k}(\underline{t})$ at continuity points of $F_{\underline{u}_k}$, and if all of the $\pi_{kn} \to \pi_k$,

then $F_{\underline{u}_n}(\underline{t}) \to F_{\underline{u}}(\underline{t}) = \sum_{k=1}^{J} F_{\underline{u}_k}(\underline{t})\,\pi_k$.

problem: If $S \not\subseteq I_K$, then

$\underline{\beta}_{K,0} \neq \underline{\beta}$ and $\sqrt{n}\,(\hat{\underline{\beta}}_{K,0} - \underline{\beta}) = \underline{u}_{Kn}$

does not converge in distribution.

However, if $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$,

then $\pi_K = 0$ for the I_K such

that $S \not\subseteq I_K$.

20] For many of the most important variable selection methods and regression methods, $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. (BIC, AIC, C_p with GLMs and MLR for p fixed).

21] Th: Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$.

$$\left.\begin{array}{lll} \text{Let } \underline{\hat{\beta}}_{MIX} = \hat{\underline{\beta}}_{I_K,0} & \text{with probs } \pi_{Kn} \\ \hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_K,0} & \text{with probs } \pi_{Kn}, \end{array}\right\} \text{same } \pi_{Kn}$$

where $\pi_{Kn} \to \pi_K$ as $n \to \infty$. Denote

The positive π_k by π_j. Assume

$$\underset{\sim}{u}_{jn} = \sqrt{n}\,(\hat{\underline{\beta}}_{I_j,0} - \underline{\beta}) \xrightarrow{D} \underline{u}_j \sim N_p(\underline{0}, V_{j,0}),$$

Assume $\underset{\sim}{w}_{jn} = \sqrt{n}(\hat{\underline{\beta}}^{C}_{I_j,0} - \underset{\sim}{\beta}) \xrightarrow{D} \underline{w}_j$.

a) $\underset{\sim}{u}_n = \sqrt{n}\,(\hat{\underline{\beta}}_{MIX} - \underset{\sim}{\beta}) \xrightarrow{D} \underline{u}$ where

the cdf $\underbrace{F_{\underline{u}}(\underline{t}) = \sum_j \pi_j \, F_{\underline{u}_j}(\underline{t})}$

$\underline{u}$ has mixture dist of the $\underset{\sim}{u}_j$

b) $\underline{w}_n = \sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta}) \xrightarrow{D} \underset{\sim}{w}$ where

$\underbrace{F_{\underset{\sim}{w}}(\underline{t}) = \sum_j \pi_j \, F_{\underset{\sim}{w}_j}(\underline{t})}$.

$\underline{w}$ has a mixture dist of the $\underset{\sim}{w}_j$

proof b) Since $\underline{w}_n$ has a mixture dist of the $\underline{w}_{kn}$ with probs π_{kn}, the cdf of $\underline{w}_n$ is $\sum \pi_{kn} F_{\underline{w}_{kn}}(\underline{t}) \rightarrow$

$F_{\underline{w}}(\underline{t}) = \sum_j \pi_j \, F_{\underline{w}_j}(\underline{t})$ at continuity points of

see 16

$F_{\underline{w}_j}(\underline{t})$ as $n \to \infty$.

Proof of a) is similar.

22] The <u>oracle property</u> is

$P(I_{min} = S) \to 1$ as $n \to \infty$.

Much stronger than $P(S \subseteq I_{min}) \to 1$

For "fast methods" like forward selection and lasso variable selection, the oracle property holds only if S is one of the J models selected with prob $\to 1$. Theory for lasso and forward selection shows the oracle property needs the predictors $x_1, \ldots, x_p$ to be "nearly orthogonal" with $\Sigma_{\underline{x}}$ diagonal. So the oracle property does not hold with correlated predictors.

23] In <u>low dimensions</u>, $\hat{\underline{\beta}}_{VS} \xrightarrow{P} \underline{\beta}_{Full}$ and $P(S \subseteq I_{min}) \to 1$ under reasonable conditions.