Suppose $\sqrt{n}(\hat{\beta}_S - \beta_S) \xrightarrow{D} N_2\left[\begin{pmatrix}0\\0\end{pmatrix}, \underbrace{\sigma^2\begin{pmatrix}w_{11} & w_{12}\\ w_{21} & w_{22}\end{pmatrix}}_{V_S}\right]$

Then $\sqrt{n}(\hat{\beta}_{S,0} - \beta) \xrightarrow{D} N_4\left[\begin{pmatrix}0\\0\\0\\0\end{pmatrix}, \sigma^2\begin{pmatrix}w_{11} & w_{12} & 0 & 0\\ w_{21} & w_{22} & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}\right]$

$\parallel$

$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_{1S}\\ \hat{\beta}_{2S}\\ 0\\ 0\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\\ 0\\ 0\end{pmatrix}\right]$.

18) $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ is a necessary condition for $\hat{\beta}_{VS}$ to be a consistent estimator of β if S is unique. The condition holds for C_p, AIC, BIC lasso VS and elastic net VS. See 17) for $u_{jn} \xrightarrow{D} u_j$.

19) $\hat{\beta}_{MIX}$ CLT: Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Let $\hat{\beta}_{MIX} = \hat{\beta}_{I_k,0}$ with probs π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $u_{jn} = \sqrt{n}(\hat{\beta}_{I_j,0} - \beta) \xrightarrow{D} u_j \sim N_p(0, V_{j,0})$.

a) Then $u_n = \sqrt{n}(\hat{\beta}_{MIX} - \beta) \xrightarrow{D} u$ where the cdf of u is $F_u(t) = \sum_j \pi_j F_{u_j}(t)$. So u has a mixture distribution of the $\underset{\sim}{u}_j$ with probs π_j, $E(\underline{u}) = \underline{0}$ (90.?)

and $Cov(\underline{u}) = \Sigma_u = \sum_j \pi_j V_{j0}$.

b) Let A be a $g \times p$ full rank constant matrix with $1 \le g \le p$. Then

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

where $\underline{v}$ has a mixture distr of the

$$\underline{v}_j = A\underline{u}_j \sim N_g(\underline{0}, A V_{j0} A^T) \text{ with probs } \pi_j.$$

c) $\hat{\underline{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\underline{\beta}$.

d) If $\pi_d = 1$, then $\sqrt{n}\,(\hat{\underline{\beta}}_{SEL} - \underline{\beta}) \xrightarrow{D} \underline{u} \sim N_p(\underline{0}, V_{d0})$

where SEL is VS or MIX.

Proof] a) Since $\underline{u}_n$ has a mixture dist of the $\underline{u}_{kn}$ with probs π_{kn}, the cdf of $\underline{u}_n$ is

$$F_{\underline{u}_n}(\underline{t}) = \sum_k \pi_{kn} F_{\underline{u}_{kn}}(\underline{t}) \to F_{\underline{u}}(\underline{t}) = \sum_j \pi_j F_{\underline{u}_j}(\underline{t})$$

at continuity points of $F_{\underline{u}}(\underline{t})$ as $n \to \infty$.

b) Since $\underline{u}_n \xrightarrow{D} \underline{u}$, then $A\underline{u}_n \xrightarrow{D} A\underline{u}$.

c) Selecting from a finite number of $\sqrt{n}$ consistent estimators (even on a set that goes to 1 in prob) results in a $\sqrt{n}$ consistent estimator.

d) If $\pi_d = 1$, there is no selection bias, asymptotically.

20) Typically the mixture distribution is not asymptotically normal.

Exceptions a) $\pi_d = 1$ (eg S = full model).

b) For each π_j, $A\underline{u}_j \sim N_g(\underline{0}, A\underset{\uparrow \text{ free of } j}{\Sigma} A^T)$

Then $\sqrt{n}\,(A\hat{\underline{\beta}}_{MIX} - A\underline{\beta}) \xrightarrow{D} N_g(\underline{0}, A\Sigma A^T)$.

b) occurs for $A\hat{\underline{\beta}}_{MIX} = \hat{\underline{\beta}}_{S,MIX}$ if

$\sqrt{n}\,(\hat{\underline{\beta}} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, V)$ where V is diagonal

and nonsingular. Then $\hat{\underline{\beta}}_{S,MIX}$ and $\hat{\underline{\beta}}_{S,FULL}$

have the same asymptotic dist, and we

Conjecture: $\hat{\underline{\beta}}_{S,VS}$ does too.

Reasoning: $Q(I) = \overbrace{(\underline{Y} - \underline{X}_I^T\hat{\underline{\beta}}_I)^T(\underline{Y} - \underline{X}_I^T\hat{\underline{\beta}}_I)}^{SSE(I)} + a\,\underset{\uparrow \text{ constant}}{c}$

if $\underline{X}_I$ is $a \times 1$, Then

$$Q(I) = (\underline{Y} - \underline{x_S^T\hat{\beta}_{S,I}} - x_{I/S}^T\hat{\beta}_{I/S})^T(\underline{Y} - \underline{x_S^T\hat{\beta}_{S,I}} - X_{I/S}^T\hat{\beta}_{I/S}) + ac$$

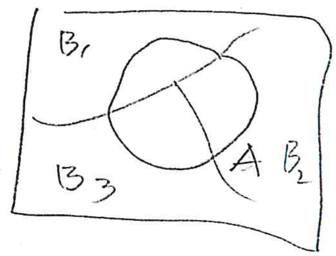
$x_S^T\hat{\beta}_{S,I}$ have the same asy dist for any I

with $S \subseteq I$. So the selection comes from $I/S \subseteq E$

which is like random selection for S.

21) $B_1, \dots, B_k$ partition S if $B_i \cap B_j = \phi$ $i \neq j$ and $P(B_i) > 0$, and $\bigcup_i B_i = S$.

5.14



Then $P(A) = \sum_{i=1}^{k} P(A \cap B_i) = \sum_{i=1}^{k} P(A|B_i)\, P(B_i)$.

Add sets $B_{k+1}, \dots, B_J$ where $P(B_j) = 0$, $j = k+1, \dots, J$

by defining $P(A|B_j) = 0$ if $P(B_j) = 0$.

Then $P(A) = \sum_{i=1}^{J} P(A|B_i)\, P(B_i)$.

22) Let $\hat{\underset{\sim}{\beta}}{}^{C}_{I_k,0} \sim \underbrace{\hat{\underset{\sim}{\beta}}_{I_k,0} \mid (\hat{\beta}_{VS} = \hat{\beta}_{I_k,0})}_{\text{conditional dist is a dist}}$

Denote $F_{\underline{z}}(\underline{t}) = P(z_1 \le t_1, \dots, z_p \le t_p)$ by $P(\underline{z} \le \underline{t})$.

Let $\underset{\sim}{w}_n = \sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta})$ and $\underline{w}_{kn} = \sqrt{n}(\hat{\underline{\beta}}{}^{C}_{I_k,0} - \underline{\beta})$.

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

$$\sum_{k=1}^{J} P\big(\overbrace{\sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta}) \le \underline{t}}^{A} \mid \overbrace{\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_k,0}}^{B_K}\big)\, P(\hat{\underline{\beta}}_{VS} = \hat{\beta}_{I_k,0})$$

(the I_k $\ni$ $P(\hat{\beta}_{VS} = \hat{\beta}_{I_k,0}) > 0$ form a partitition)

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

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

So $\hat{\underline{\beta}}_{VS}$ has a mixture dist of the $\hat{\underline{\beta}}_{I_k,0}$ with probs π_{kn}

and " $\underline{w}_n$ $\qquad\qquad$ $\underline{w}_{kn}$ "

are selecting from $\underline{U}_{kn} = \sqrt{n}\,(\hat{\underline{\beta}}_{I_k,0} - \underline{\beta})$ and asymptotically from $\underline{U}_k$ (where $\underline{U}_k \sim N_p(\underline{0}, V_{k0})$) if $S \subseteq I_k$).
Random selection does not change the dist of $\underline{U}_{kn}$ and $\underline{U}_k$, but variable selection changes the dist of selected $\underline{U}_{kn}$ to $\underline{W}_{kn}$ and selection bias changes the dist of $\underline{U}_j$ to $\underline{W}_j$.

24) <u>Variable selection CLT!</u> Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ and let $\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_k,0}$ with probs π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\underline{W}_{jn} = \sqrt{n}\,(\hat{\underline{\beta}}^C_{I_j,0} - \underline{\beta}) \xrightarrow{D} \underline{W}_j$. Then $\underline{W}_n = \sqrt{n}\,(\hat{\underline{\beta}}_{VS} - \underline{\beta}) \xrightarrow{D} \underline{W}$ where the cdf of $\underline{W}$ is $F_{\underline{W}}(\underline{t}) = \sum_j \pi_j F_{\underline{W}_j}(\underline{t})$. Thus $\underline{W}$ is a mixture dist of the $\underline{W}_j$ with prob's π_j.

proof) By 22, $F_{\underline{W}_n}(\underline{t}) \overset{\text{by (22)}}{=} \sum_k \pi_{kn} F_{\underline{W}_{kn}}(\underline{t}) \xrightarrow{D} F_{\underline{W}}(\underline{t}) = \sum_j \pi_j F_{\underline{W}_j}(\underline{t})$ at continuity points of the $F_{\underline{W}_j}(\underline{t})$ as $n \to \infty$.

25) $\hat{\underline{\beta}}_{S,MIX}$ seems to be a good approx for $\hat{\underline{\beta}}_{S,VS}$ unless the predictors are highly correlated. Most of the selection bias is due to predictors in E which make selection of S almost random. $\hat{\underline{\beta}}_{E,MIX}$ and $\hat{\underline{\beta}}_{E,VS}$ tend to differ, but both use zero padding.
If $\pi_d = 1$ and $S \subseteq I_d$ then $\sqrt{n}\,(\hat{\underline{\beta}}_{SEL} - \underline{\beta}) \xrightarrow{D} N_p(0, V_{d0})$, SEL = MIX or VS.

26) Suppose $T_i^{*} = T_{ij}$ with prob δ_{jn} for $j=1,\dots,J$.

Then the bootstrap sample $T_1^*, \dots, T_B^* =$

52.5

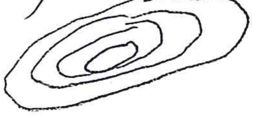
$$\underbrace{T_{1,1}^*, \dots, T_{B_{1n},1}^*}_{\text{1st bootstrap component}}, \dots, \underbrace{T_{1,J}^*, \dots, T_{B_{Jn},J}^*}_{J\text{th}}.$$

Denote $T_{1j}^*, \dots, T_{B_{jn},j}^*$ as the jth bootstrap component with sample mean $\overline{T}_j^*$ and sample covariance matrix $S_{T,j}^*$. Define the jth component of the iid sample $T_1, \dots, T_B$ to have sample mean $\overline{T}_j$ and sample covariance matrix $S_{T,j}$

27) Under regularity conditions, if $S \subseteq I_j$ then $\sqrt{n}(\hat{\beta}_{I_j,0} - \beta) \xrightarrow{D} N_p(0, V_{j0})$ and $\sqrt{n}(\hat{\beta}^*_{I_j,0} - \hat{\beta}_{I_j,0}) \xrightarrow{D} N_p(0, V_{j0})$.

Then the component clouds have the same variability asymptotically. The iid data component clouds are centered at β and the geometric argument holds for the iid data cloud and $\hat{\beta}_{MIX}$. If the bootstrap data clouds were centered at $\tilde{\beta}$, the bootstrap conf regions would work.

Instead, the bootstrap data clouds are shifted slightly from a common center $\overline{T}^*$ and are each centered at $\hat{\beta}_{I_j,0}$. Geometrically, the shifting of the bootstrap data clouds makes the bootstrap data cloud more variable than the iid data cloud asymptotically, resulting in slightly higher asymptotic coverage.



$\leftarrow$ iid data clouds have common center

585 53

28) For $\hat{\beta}_{VS}$ the bootstrap component clouds are also shifted.

29) $A\hat{\beta}_{MIX} = T_n$ is not observed so the hybrid and Bickel and Ren regions plugged in $A\hat{\beta}_{VS}$ instead.

30) For $n \geq 20p$ and $B \geq \max(200, 50p)$ (so S_T^* is a good estimator of $Cov(T^*)$), simulations for $\hat{\beta}_{VS}$ were ~~often~~ better than those for $\hat{\beta}_{MIX}$.

31) S_T^* can be singular due to one or more columns of zeroes in the bootstrap sample for $\beta_1, \ldots, \beta_p$

eg if X_k is never selected, the kth column is all zeroes. Then X_k is likely not needed in the MLR model.

Remedy add $d = \lceil 0.01 B \rceil$ bootstrap samples of the full model.

32) under coverage can occur if $\frac{n-p}{n}$ is not close to 1.

coverage can be higher than the nominal coverage if i) the bootstrap data cloud is more variable than the iid data cloud ii) zero padding.

ex] X_k never selected → shorth CI $= [0, 0]$ which has 100% coverage if $\beta_k = 0$.

33) parametric bootstrap: (5.5.9)

$$\underset{\sim}{Y}^* \sim \underbrace{N_n(X\hat{\underline{\beta}}, \overbrace{MSE}^{\text{full model}}\, I_n)}_{\text{known parametric dist}} \sim N_n(H\underset{\sim}{Y}, MSE\, I_n)$$

so $\underline{Y}^* = X\hat{\underline{\beta}} + \underline{e}^*$ with $e_i^* \overset{iid}{\sim} N(0, MSE)$

Then $\hat{\underline{\beta}}_I^* = (X_I^T X_I)^{-1} X_I^T \underline{Y}^* \sim N_{a_I}[\hat{\underline{\beta}}_I, MSE(X_I^T X_I)^{-1}]$

since $E(\hat{\underline{\beta}}_I^*) = (X_I^T X_I)^{-1} X_I^T H Y = \hat{\underline{\beta}}_I$ since

$H X_I = X_I$, and $Cov(\hat{\underline{\beta}}_I^*) = MSE\, (X_I^T X_I)^{-1}$.

Thus $\sqrt{n}(\hat{\underline{\beta}}_I^* - \hat{\underline{\beta}}_I) \sim N_{a_I}(\underline{0}, n\, MSE\, (X_I^T X_I)^{-1})$

$\xrightarrow{D} N_{a_I}(\underline{0}, V_I)$ as $n, B \to \infty$ if $S \subseteq I$.

34) residual bootstrap:

$\underline{Y}^* = X\hat{\underline{\beta}} + \underline{r}^w$ where $\underline{e}^* = \underline{r}^w$ is a $n \times 1$ vector of elements selected with replacement from the OLS full model residuals. $E(r_i^w) = \frac{1}{n}\sum_{i=1}^n r_i = 0$, $V(r_i^w) = E[(r_i^w)^2]$

$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^n r_i^2 = \frac{n-p}{n} MSE$, $E(\underline{r}^w) = \underline{0}$ $Cov(\underline{r}^w) = \sigma_n^2 I_p$

$\sqrt{n}(\hat{\underline{\beta}}^* - \hat{\underline{\beta}}) \xrightarrow{D} N_p(\underline{0}, \sigma^2 V)$.

$E(\hat{\underline{\beta}}_I^*) = \hat{\underline{\beta}}_I$ $\quad Cov(\hat{\underline{\beta}}_I^*) \approx Cov(\hat{\underline{\beta}}_I)$,

Conjecture $\sqrt{n}(\hat{\underline{\beta}}_I^* - \hat{\underline{\beta}}_I) \xrightarrow{D} N_{a_I}(0, V_I)$.

$\Bigg\} S \subseteq I$

35} Nonparametric bootstrap

select n cases with replacement from $(y_1, \underline{x}_1), \ldots, (y_n, x_n)$ to form $\underline{Y}^*$, $\underline{X}^*$.
$(y_1, \underline{x}_1, r_1)$ (y_n, x_n, r_n)

Then $\underline{Y}^* = \underline{X}^* \hat{\underline{\beta}} + \underline{r}^W$ and

$$\underline{Y}^* = X_I^* \hat{\beta}_I + \underline{r}_I^W.$$ Under regularity conditions,

$\sqrt{n} \, (\hat{\underline{\beta}}^* - \hat{\underline{\beta}}) \xrightarrow{D} N_p(0, V)$ and

$\sqrt{n} \, (\hat{\beta}_I^* - \hat{\beta}_I) \xrightarrow{D} N_p(0, V_I)$ if $S \subseteq I$.

§7.5

36} Data splitting: randomly select n_H cases for H and $n - n_H = n_V$ cases for V.

Build model and do variable selection for cases in H to get model I_{min}.

Fit I_{min} to cases in V and use usual inference.

want want $10p \le n_H \le \frac{n}{2}$ with $n_H \approx 10p$.

§7.6

37} There are several important alternatives to OLS for MLR.

Need ... the method ... using

the same measurements gets the same $(\hat{\underline{Y}}, \hat{\underline{\beta}})$. (4.5)

Let $\underline{x}_i = (1, \underline{u}_i^T)$ where $\underline{u}_i = (x_{i2}, \ldots, x_{ip})^T$

are the nontrivial predictors. Let

$\underline{w}_i$ be the standardized nontrivial predictors

so $\overline{\underline{w}} = \underline{0}$ and $\frac{1}{n}\sum_{i=1}^{n} w_{ij}^2 = 1$.

Let $\underline{Z} = \underline{Y} - \overline{\underline{Y}}$ where $\overline{\underline{Y}} = \overline{Y}\,\underline{1}$. Do regression

through the origin for $\underline{Z} = W \underline{\eta} + \underline{e}$ where

$\underline{\eta} = (\eta_1, \ldots, \eta_{p-1})^T$. Then $\hat{\underline{Y}} = \overline{\underline{Y}} + \hat{\underline{Z}}$ and

$\hat{\underline{\beta}}$ can be found from $\hat{\underline{\eta}}$.

ex) <u>Know for Quiz 9</u> $\underline{Y} = X\underline{\beta} + \underline{e}$ and

MLR method fits $\underline{Z} = W\underline{\eta} + \underline{e}$

suppose $\hat{Z} = 245.3$ and $\overline{Y} = 105.37$, what is $\hat{Y}$?

Soln) $\hat{Y} = \hat{Z} + \overline{Y} = \boxed{351.00}$

38) Get a PI $[r_L, r_U]$ for the residuals.

Then a PI for Y_f is $[\hat{Y}_f + r_L, \hat{Y}_f + r_U]$.

ex) <u>Know for Quiz 9</u>: 90% PI for r_f is $[-778.28, 1336.44]$

$\hat{\underline{\beta}}_{Imin} = (241.545, 1.001)^T$, $\underline{x}_f = (1, 75000)^T$, $\hat{Y}_f = \underline{x}_f^T \hat{\underline{\beta}}_{Imin} =$

$241.545 + 1.001(75000) = 75316.545$. So $[74538.265, 76652.985]$

is a 90% P.I. for Y_f.