confidence region is $\{\underline{0}\}$, $n \geq 20p$, $B \geq 50p$ and the error dist is unimodal and not highly skewed. (This technique may be useful for data snooping.)

ex) $\sigma^2 V = \sigma^2 I$ $\quad Y = \beta_1 + \beta_2 x_2 + \varepsilon = 1 + x_2 + \varepsilon$

$= 1 + x_2 + 0x_3 + 0x_4 + \varepsilon$ so $p=4$.

$\beta_S = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ $\quad \beta_E = \begin{pmatrix} \beta_3 \\ \beta_4 \end{pmatrix}$.

Test H_0 $\underline{\beta_S} = (1)$ with PRI BRI hybl

H_0 $\beta_E = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ PRO BRO hypo.

Nominal 95% 5000 runs. $B=1000$, $n=100=25p$

Apply shorth CI for $\hat{\beta}_{i1}^*, \ldots \hat{\beta}_{iB}^*$ slight undercoverage

	$\beta_1=1$	$\beta_2=1$	$\beta_3=0$	$\beta_4=0$	$H_0 \beta_E = \underline{0}$ PRO	hypo	BRO	$H_0 \underline{\beta_S} = \underline{1}$ PRI	hybl	BRI
reg	.946	.950	.947	.948	.940	.941	.941	.937	.936	.937
len	.396	.399	.394	.398	2.451	2.451	2.452	2.450	2.450	2.451
vs	.948	.950	.997	.996	.991	.974	.991	.938	.939	.940
len	.395	.398	.323	.323	2.694	2.699	3.002	2.450	2.450	2.457

vs and full similar

vs coverage higher, volume likely lower

vs and full similar

Zero padding makes coverage higher but volume (length CI) smaller

$D_{UB,I} = D_{UB}$

For β_1, β_2 pop CI length $= 0.392$

For confidence region "length" is D_{UB}^2 or $D_{UB,I}^2$ the cutoff. $\sqrt{\chi^2_{2,.95}} = 2.45$

can't compare vs vs reg since vol is not given.

43) Many regression models satisfy
$Y \perp\!\!\!\perp \underline{x} \mid \underline{x}^T\underline{\beta}$ with $\sqrt{n}(\hat{\underline{\beta}} - \underline{\beta}) \xrightarrow{D} N_p(0, W)$.

eg GLMs, survival regression, MLR.
For variable selection with AIC
nearly the same argument as that
for MLR works with parametric
bootstrap.
if $Y|\underline{x} \sim D(\underline{\beta}^T\underline{x}, \underline{\theta})$.

Poisson reg $Y|\underline{x} \sim \text{Poisson}(e^{\underline{\beta}^T\underline{x}})$
bin reg $Y_i|x_i \sim \text{bin}\left(m_i, \frac{e^{\underline{\beta}^T\underline{x}_i}}{1+e^{\underline{\beta}^T\underline{x}_i}}\right)$.
multivariate regression for survival analysis
GLMs etc.

44) Large sample PI for Y_t given $\underline{x}_t$
Apply shorth large sample $(100(1-\delta)\%$ PI
to $Y_1^*, \ldots, Y_B^*$ i.i.d $D(\hat{\underline{\beta}}^T\underline{x}_t, \hat{\theta})$
obtained from the regression

close to 1.

coverage can be higher than the nominal coverage for 2 reasons:

i) the bootstrap data cloud is more variable than the iid data cloud

ii) zero padding $\hat{\beta}_{I_{min}}, 0$.

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

where V is diagonal,

$\sqrt{n}(\hat{\beta}_S - \beta_S) \xrightarrow{D} N_{a_S}(0, \sigma^2 V_S)$

for both the full model $\hat{\beta}$ and $\hat{\beta}_{MIX}$ and maybe for $\hat{\beta}_{VS}$.

If $\theta = A\beta_S$ expect similar coverage and volume for full model and variable selection confidence regions.

46) Let $\beta^T x = \beta_S^T x_S + \beta_E^T x_E = \beta_S^T x_S$.

Let H_0 $A\beta = \beta_O = 0$ where $\beta_O = (\beta_{i_1}, \ldots, \beta_{i_g})^T$, $O \subseteq E$ so H_0 is true, β_I = other $p-g$ terms

a) Suppose a nominal 95% confidence region is used and $V_B = 0.96B$.

Then the prediction region and Bickel and Ren confidence regions contain at least 96% of the bootstrap sample. If $\hat{\beta}^*_{0,j} = 0$ for more than 4% of the $\hat{\beta}^*_{0,1}, \ldots, \hat{\beta}^*_{0,B}$ then 0 is in the confidence region and the bootstrap test fails to reject H_0. If this occurs for each run of the simulation (5000 runs), then the observed coverage is 100% > 95%.

b) Now suppose $\hat{\beta}^*_{0,j} = 0$ for $j = 1, \ldots, B$. Then S^*_j is singular but $\{0\}$ is the large sample $100(1-\delta)\%$ confidence region for each of the 3 bootstrap methods: Prediction region, Bickel and Ren, hybrid.

assume $\beta_0 = \ln 2$ for forward selection

Then pval = 1 estimates the pop pvalue.

For the Imin Model for forward selection, there may be strong evidence that x_0 is not needed in the model given x_I is in the model (if the "100%"

d12.5 1) Lasso, relaxed lasso, vs

Stc.P to mreg

ridge regression & elastic net are other ways to fit $\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$. They are especially useful if $X'X$ is singular, or if $(X'X)^{-1}$ is ill conditioned, eg if $p >> n$.

2) Lasso, relaxed lasso vs and elastic net do variable selection, some $\hat{\beta}_i = 0$ are possible.

3) Let the nontrivial predictors $\underset{\sim}{u}_i = (x_{i2}, \ldots, x_{ip})^T$. Let the $n \times (p-1)$ matrix of standardized nontrivial predictors be $W = (w_{ij})$ where

$\sum_{i=1}^{n} w_{ij} = 0$ and $\sum_{i=1}^{n} w_{ij}^2 = n$

jth standardized predictor has $\overline{w_j} = 0$

$\tilde{\sigma}_j^2 = \frac{1}{n} \sum_{i=1}^{n} w_{ij}^2 = 1$

Then the sample correlation matrix of the nontrivial predictors is

$$R_{\underset{\sim}{u}} = \frac{W^T W}{n}. \quad \text{Let } \underset{\sim}{Z} = \underset{\sim}{Y} - \overline{\underset{\sim}{Y}}$$

where $\overline{\underset{\sim}{Y}} = \overline{Y} \underset{\sim}{1}$. The MLR methods in 1) fit $\underset{\sim}{Z} = W\underset{\sim}{\eta} + \underset{\sim}{e}$ and then find $\hat{\underset{\sim}{\beta}}$ from $\hat{\underset{\sim}{\eta}}$ and $\hat{\underset{\sim}{Y}} = \hat{\underset{\sim}{Z}} + \overline{\underset{\sim}{Y}}$.

4) Let $Q(\underline{\eta}) = \frac{1}{a}(\underline{z} - \underline{w}\underline{\eta})^T(\underline{z} - \underline{w}\underline{\eta}) + \frac{\lambda_n}{a}\sum_{i=1}^{p-1} |\eta_i|^j$

where $\lambda_n \geq 0$, $a > 0$ and $j > 0$ are known.

Then $j=2$ gives ridge regression, $j=1$ gives lasso and $a = 1, 2, \eta$, and $2n$ are common.

The residual sum of squares $RSS(\underline{\eta}) = (\underline{z} - \underline{w}\underline{\eta})^T(\underline{z} - \underline{w}\underline{\eta})$, and

$\lambda_n = 0$ corresponds to the OLS estimator $\underline{\hat{\eta}}_{OLS} = (\underline{w}^T\underline{w})^{-1}\underline{w}^T\underline{z}$

5) Lasso and ridge regression use a grid of m λ values $0 \leq \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{m-1} \leq \lambda_m$.
For lasso, λ_m is the smallest value of λ such that $\underline{\hat{\eta}}_{\lambda_m} = \underline{0}$.

Hence $\underline{\hat{\eta}}_{\lambda_i} \neq \underline{0}$ for $i < m$.

6) The elastic net estimator minimizes the criterion

$Q(\underline{\eta}) = RSS(\underline{\eta}) + \lambda_1 \|\underline{\eta}\|_2^2 + \lambda_2 \|\underline{\eta}\|_1$
where $\lambda_1 = (1-\alpha)\lambda_n$, and $\lambda_2 = 2\alpha\lambda_n$ with $0 \leq \alpha \leq 1$.

7) Assume $(W^TW)^{-1}$ exists.

These estimators minimize convex Q, Karush-Kuhn-Tucker (KKT) conditions for convex optimality exist and are like normal equations.

KKT for lasso: $-\frac{1}{n} W^T(\underline{z} - W\hat{\underline{\eta}}_L) + \frac{\lambda_1 n}{2n} \underline{s}_n = \underline{0}$

or $-W^T(\underline{z} - W\hat{\underline{\eta}}_L) + \frac{\lambda_1 n}{2} \underline{s}_n = \underline{0}$

where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_{iL})$ if $\hat{\eta}_{iL} \neq 0$. $\text{sign}(\eta_i) \begin{cases} 1 & \eta_i > 0 \\ -1 & \eta_i < 0 \end{cases}$

So $\underline{s}_n = s_n(\hat{\underline{\eta}}_L)$ depends on $\hat{\underline{\eta}}_L$. Thus

$$\hat{\underline{\eta}}_L = (W^TW)^{-1}W^T\underline{z} - \frac{\lambda_1 n}{2n} n (W^TW)^{-1}\underline{s}_n =$$

$$\hat{\underline{\eta}}_{OLS} - \frac{\lambda_1 n}{2n} n (W^TW)^{-1}\underline{s}_n$$

KKT for EN: $2W^TW\hat{\underline{\eta}}_{EN} - 2W^T\underline{z} + 2\lambda_1\hat{\underline{\eta}}_{EN} + \lambda_2\underline{s}_n = \underline{0}$ After algebra,

$$\hat{\underline{\eta}}_{EN} = \hat{\underline{\eta}}_{OLS} - n\left(W^TW + \lambda_1 I_{(p-1)}\right)^{-1}\left[\frac{\lambda_1}{n}\hat{\underline{\eta}}_{OLS} + \frac{\lambda_2}{2n}\underline{s}_n\right]$$

a) If $\frac{\lambda_1 n}{\sqrt{n}} \xrightarrow{p} 0$, then $\sqrt{n}(\hat{\underline{\eta}}_A - \underline{\eta}) \xrightarrow{D} N_{p-1}(0, \sigma^2 V)$

(probably not tekT)

identity of RR: $\hat{\eta}_R = (W^TW + \lambda_{1n} I_{p-1})^{-1} W^T Z$

$= (W^TW + \lambda_{1n} I_{p-1})^{-1} W^TW (W^TW)^{-1} W^T Z$

$= (W^TW + \lambda_{1n} I_{p-1})^{-1} W^TW \hat{\eta}_{OLS} = A_n \hat{\eta}_{OLS} =$

$[I_{p-1} - \lambda_{1n} (W^TW + \lambda_{1n} I_{p-1})^{-1}] \hat{\eta}_{OLS} =$

$B_n \hat{\eta}_{OLS} = \hat{\eta}_{OLS} - \frac{\lambda_{1n}}{n} (W^TW + \lambda_{1n} I_{p-1})^{-1} \hat{\eta}_{as}$

Since $A_n - B_n = 0$.

8) Assume the sample correlation matrix

(*) $R_U = \frac{W^TW}{n} \xrightarrow{P} V^{-1}$ where

$V^{-1} = \rho_U$ = the pop correlation matrix

of the nontrivial predictors U_i are iid from a pop.

If $\frac{\lambda_{1n}}{n} \to 0$, then $\frac{W^TW + \lambda_{1n} I_{p-1}}{n} \xrightarrow{P} V^{-1}$

and $n(W^TW + \lambda_{1n} I_{p-1})^{-1} \xrightarrow{P} V$.

9) R) Let $\hat{\eta}_A$ be $\hat{\eta}_{EN}$, $\hat{\eta}_L$ or $\hat{\eta}_R$.

Assume $\sqrt{n}(\hat{\eta}_{OLS} - \eta) \xrightarrow{D} N_{p-1}(0, \sigma^2 V)$.

a) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} 0$, $\sqrt{n}(\hat{\underline{\eta}} - \underline{\eta}) \xrightarrow{D} N_{p-1}(\underline{0}, \sigma^2 V)$

b) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$ and $\underline{s}_n \to \underline{s} = \underline{s}_\eta$

then $\sqrt{n}(\hat{\underline{\eta}}_{EN} - \underline{\eta}) \xrightarrow{D} N_{p-1}(-V[(1-\psi)\tau\underline{\eta} + \psi\tau\underline{s}], \sigma^2 V)$.

c) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$, then

$\sqrt{n}(\hat{\underline{\eta}}_R - \underline{\eta}) \xrightarrow{D} N_{p-1}(-\tau V \underline{\eta}, \sigma^2 V)$.

d) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$ and $\underline{s}_n \xrightarrow{P} \underline{s} = \underline{s}_\eta$,

then $\sqrt{n}(\hat{\underline{\eta}}_L - \underline{\eta}) \xrightarrow{D} N_{p-1}(\frac{\tau}{2} V \underline{s}, \sigma^2 V)$.

10) Lasso and ridge regression are consistent estimators of $\underline{\eta}$ if $\frac{\lambda_n}{n} \to 0$ as $n \to \infty$,

$\sqrt{n}$ consistent if $\frac{\lambda_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$

and asymptotically equivalent to OLS if $\frac{\lambda_n}{\sqrt{n}} \xrightarrow{P} 0$.

11) Roughly lasso sets about one $\hat{\beta}_i = 0$ if $\frac{\lambda_n}{\sqrt{n}} \xrightarrow{P} \tau > 0$.

12) Relaxed lasso is OLS applied to the x_i that have lasso $\hat{\beta}_i \neq 0$.

Let I_{min} be this set. If $P(S \subseteq I_{min}) \to 1$, then relaxed lasso theory is like that of forward selection except the π_j differ.

Relaxed elastic net is OLS applied to the x_i that have EN $\hat{\beta}_i \neq 0$. Expect relaxed versions perform better unless $(X_{I_{min}}' X_{I_{min}})^{-1}$ is ill conditioned.

(13) Proof Sketch for (9)

If $\frac{\hat{\lambda}_{1n}}{\sqrt{n}} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then

$\frac{\hat{\lambda}_1}{\sqrt{n}} \xrightarrow{P} (1-\psi)/\tau$ and $\frac{\hat{\lambda}_2}{\sqrt{n}} \xrightarrow{P} 2\psi\tau$. Then

$$\sqrt{n}(\hat{\eta}_{EN} - \eta) = \underbrace{\sqrt{n}(\hat{\eta}_{OLS} - \eta)}_{\xrightarrow{D} N_{p-1}(0, \sigma^2 V)} - \underbrace{n(W^T W + \hat{\lambda}_{1n} I_{p-1})}_{\xrightarrow{P} -V} \underbrace{\left[\frac{\hat{\lambda}_{1n}}{\sqrt{n}} \hat{\eta}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}} s_\eta\right]}_{\xrightarrow{P} (1-\psi)/\tau \eta + 4\tau s}$$

$$\text{so } \sqrt{n}(\hat{\eta}_{EN} - \eta) \xrightarrow{D} N_{p-1}(0, \sigma^2 V) - V[(1-\psi)/\tau \eta + 4\tau s]$$

$$\sim N_{p-1}\left(\underbrace{-V[(1-\psi)/\tau \eta + 4\psi\tau s]}_{0 \text{ if } \tau = 0}, \sigma^2 V\right)$$

14) $\hat{\underline{n}}_R = (\underline{W}^T\underline{W} + \lambda_{in} I_{p+1})^{-1} \underline{W}^T \underline{z}$.

$\underline{W}^T\underline{W}$ is symmetric and sgcare

so $\underline{W}^T\underline{W} \geq 0$ with eigenvalues $\psi_1 \geq \cdots \geq \psi_p \geq 0$, If $\psi_p = 0$ then $(\underline{W}^T\underline{W})^{-1}$ does not exist.

Let $(\psi, \underline{g})$ be an eigenvector eigenvector pair of $\underline{W}^T\underline{W} = n R_U$.

Then $[\underline{W}^T\underline{W} + \lambda_{in} I_{p+1}]\underline{g} =$

$\underline{W}^T\underline{W}\,\underline{g} + \lambda_{in}\underline{g} = \psi\underline{g} + \lambda_{in}\underline{g} =$

$(\psi + \lambda_{in})\underline{g}$. So $(\underbrace{\psi + \lambda_{in}}_{>0 \text{ if } \psi_{in} > 0}, \underline{g})$

is an eigenvalue eigenvector pair of $\underbrace{\underline{W}^T\underline{W} + \lambda_{in} I_{p+1} > 0}_{\text{positive definite}}$ if $\lambda_{in} > 0$.

Hence $(\underline{W}^T\underline{W} + \lambda_{in} I_{p+1})^{-1}$ exists $\forall\, \lambda_{in} > 0$, even if $\underline{W}^T\underline{W}$ is singular or ill conditioned.

12) Let $A = [\underset{\sim}{a}_1 \cdots \underset{\sim}{a}_K]$. Then the vec operator stacks the columns of A so

$$\text{vec}(A) = \begin{pmatrix} \underset{\sim}{a}_1 \\ \underset{\sim}{a}_2 \\ \vdots \\ \underset{\sim}{a}_K \end{pmatrix}.$$

13) Let $\underset{m \times n}{A} = (a_{ij})$ and B be $p \times q$. The Kronecker product of A and B is

$$\underbrace{A \otimes B}_{mp \times nq} = \begin{pmatrix} a_{11}B & a_{12}B & \cdots & a_{1n}B \\ a_{21}B & a_{22}B & \cdots & a_{2n}B \\ \vdots & & & \\ a_{m1}B & a_{m2}B & \cdots & a_{mn}B \end{pmatrix}.$$

If A and B are non singular, then $[A \otimes B]^{-1} = A^{-1} \otimes B^{-1}$. $A \otimes B$ is a compact method for writing down certain block matrices.

14) The multivariate least squares central limit theorem (MLS CLT). Suppose $\max(h_1, \ldots, h_n) \to 0$, $\frac{1}{n} X^T X \xrightarrow{P} W^{-1}$ and the zero mean i.i.d $\underset{\sim}{\varepsilon}_i$ have finite 4th moments,

Then $\sqrt{n}\,\mathrm{vec}(\hat{B}-B) \xrightarrow{D} N_{pm}\left(\underset{\sim}{0}, \underset{m\times m}{\Sigma_\varepsilon} \otimes W\right)$.

Note: LS CLT had the same assumptions except the zero mean iid ε_i had 2nd moments, 4th moments are likely used because a bigger theorem also showed $\hat{\Sigma}_\varepsilon$ is $\sqrt{n}$ consistent and asymptotically normal. If $m=1$, $\Sigma_\varepsilon = \sigma^2$ and $\Sigma_\varepsilon \otimes W = \sigma^2 W$.

15) To test $H_0: LB=0$ vs $H_1: LB\neq 0$, let

$W_e = \hat{E}^T\hat{E} = (n-p)\hat{\Sigma}_\varepsilon$, let

$H = \hat{B}^T L^T \left[L(X^TX)^{-1}L^T\right]^{-1} L\hat{B}$, and let

$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the eigenvalues of $W_e^{-1}H$.

Here H is not the hat matrix.

Wilks' $\Lambda(L) = \prod_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$,

Pillai's $V(L) = \sum_{i=1}^{m} \frac{\lambda_i}{1+\lambda_i}$, and

Hotelling Lawley $U(L) = \sum_{i=1}^{m} \lambda_i =$

$$\frac{1}{n-p}\left[\mathrm{vec}(L\hat{B})\right]^T\left[\hat{\Sigma}_\varepsilon^{-1}\otimes\left(L(X^TX)^{-1}L^T\right)^{-1}\right]\mathrm{vec}(L\hat{B}).$$

16) Under regularity conditions, if H_0 is true

$$-\left[n-p+1-0.5(m-r+3)\right]\log\left[\Lambda(L)\right]\xrightarrow{D}\chi^2_{rm}$$

$$(n-p)\,V(L)\xrightarrow{D}\chi^2_{rm} \quad \text{and}$$

$$(n-p)\,U(L)\xrightarrow{D}\chi^2_{rm}.$$

17) For $m=1$, the Seber test statistic for $H_0: L\underset{\sim}{\beta}=\underset{\sim}{0}$ was

$$F_R=\frac{\left[L\hat{\beta}\right]^T\left[L(X^TX)^{-1}L^T\right]^{-1}\left[L\hat{\beta}\right]}{r\hat{\sigma}^2}$$

$$=\frac{(n-p)\,U(L)}{r} \quad \text{since } \hat{\Sigma}_\varepsilon^{-1}=\frac{1}{\hat{\sigma}^2}.$$

So the Hotelling Lawley test statistic is the MLR partial F test statistic extended to $m>1$ response variables.

18) use $F_{rm,\,n-rm}$ approx instead of χ^2_{rm}!

$$pval = P\left[F_{rm,n-rm} > \frac{-[n - p + 1 - 0.5(m - r + 3)][\log(|L|)]}{rm} \right]$$

$$pval = P\left[F_{rm,n-rm} > \frac{(n-p)\, V(L)}{rm} \right]$$

$$pval = P\left[F_{rm,n-rm} > \frac{(n-p)\, U(L)}{rm} \right].$$

These are large sample tests.

19) The Mahalanobis distance of a $p \times 1$ vector $\underset{\sim}{w}_i$ with respect to a multivariate location and dispersion estimator $(\underset{p\times 1}{T}, \underset{p\times p}{C})$ is

$$D_i(T,C) = D_{\underset{\sim}{w}_i}(T,C) = \sqrt{(\underset{\sim}{w}_i - T)^T C^{-1} (\underset{\sim}{w}_i - T)}.$$

The classical Mahalanobis distance uses the sample mean

$(T,C) = (\overline{\underset{\sim}{w}}, S_{\underset{\sim}{w}})$ where $\overline{\underset{\sim}{w}} = \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{w}_i$ and the sample covariance matrix estimator

$$S_{\underset{\sim}{w}} = \frac{1}{n-1} \sum_{i=1}^{n} (\underset{\sim}{w}_i - \overline{\underset{\sim}{w}})(\underset{\sim}{w}_i - \overline{\underset{\sim}{w}})^T.$$

20) Let $MD_i = D_i(\overline{\underset{\sim}{w}}, S_{\underset{\sim}{w}})$ and let $RD_i = D_i(T,C)$