$$\frac{\underset{\sim}{\wedge}_1^T \underline{\Sigma}_{xy}}{\underbrace{\underset{\sim}{\wedge}_1^T \underline{\Sigma}_x \underset{\sim}{\wedge}_1^T}_{\lambda,\ \text{scalar}}} \underset{p\times 1}{\underset{\sim}{\wedge}_1} = \underset{\sim}{\wedge}_q \left(\underset{\sim}{\wedge}_q^T \underline{\Sigma}_x \underset{\sim}{\wedge}_q\right)^{-1} \underset{\sim}{\wedge}_q^T \underline{\Sigma}_{xy} = \underline{\Sigma}_x^{-1} \underline{\Sigma}_{xy}$$

for $q = 1, \ldots, p$.

PLS theory shows $\underset{\sim}{\wedge}_1 = \underline{\Sigma}_{xy}$.

Hence $\lambda \underline{\Sigma}_{xy} = \underline{\Sigma}_x^{-1} \underline{\Sigma}_{xy} = \underline{\beta}$.

eigenvalue ↑ ↑ eigenvector.

17) There are $p+1$ PLS estimators

$\left(\hat{\alpha}_{q,PLS}, \underline{\hat{\beta}}_{q,PLS}\right)$ for $q = 1, \ldots, p$ where

$$\left(\hat{\alpha}_{p,PLS}\ \underline{\hat{\beta}}_{p,PLS}\right) = \left(\hat{\alpha}_{OLS}, \underline{\hat{\beta}}_{OLS}\right)$$

and $\left(\hat{\alpha}_{q^*,PLS}\ \underline{\hat{\beta}}_{q^*,PLS}\right)$ where q^* is found from a model selection technique, such as K-fold cross validation.

Interest is in the one component PLS estimator $\underline{\hat{\beta}}_{1,PLS} = \hat{\lambda}\, \underline{\hat{\Sigma}}_{xy} = \hat{\lambda}\, \underline{\hat{\alpha}}$.

$$\hat{\lambda} = \frac{\hat{\underline{\Sigma}}_{xy}^T \hat{\underline{\Sigma}}_{xy}}{\hat{\underline{\Sigma}}_{xy}^T \hat{\Sigma}_x \hat{\underline{\Sigma}}_{xy}}$$

. Compute

$\hat{\underline{\eta}} = \hat{\underline{\Sigma}}_{xy} = \widehat{cov}(\underline{X}, Y)$. Then compute

$\underline{w}_i = \hat{\underline{\eta}}^T \underline{X}_i$ for $i = 1, \ldots, n$.

Fit the working simple linear regression model $Y_i = \alpha + \lambda w_i + \varepsilon_i$ $i = 1, \ldots, n$ with OLS to get $\hat{\alpha}_{IPLS}$ and $\hat{\lambda}$.

18) Let $\underline{\Sigma}_{\underline{w}} = \underline{\Sigma}_{\underline{\eta}}$. Let $\hat{\underline{\beta}}_{OPLS} = \hat{\underline{\beta}}_{IPLS}$.

19) Th: Assume

$$\sqrt{n}\left[\begin{pmatrix} \hat{\lambda} \\ \hat{\underline{\eta}} \end{pmatrix} - \begin{pmatrix} \lambda \\ \underline{\eta} \end{pmatrix}\right] \xrightarrow{D} N_{p+1}\left[\begin{pmatrix} 0 \\ \underline{0} \end{pmatrix}, \begin{pmatrix} \sigma_\lambda & \underline{\sigma}_{\lambda\eta} \\ \underline{\sigma}_{\eta\lambda} & \underline{\Sigma}_\eta \end{pmatrix}\right]$$

(scalar)

$$\sim N_{p+1}(\underline{0}, \underline{\Sigma}).$$

a) $\sqrt{n}(\hat{\underline{\mu}} - \underline{\mu}) \xrightarrow{D} N_p(\underline{0}, \Sigma_1) - N_p(\underline{0}, \Sigma_w)$

by LS theory for $\hat{\underline{\mu}} = \bar{x}_y$ (large sample) given in ex on notes 77].

See Exam 3 review.

b) $\sqrt{n}(\hat{\lambda}\hat{\underline{\mu}} - \lambda\underline{\mu}) = \sqrt{n}(\hat{\underline{B}}_{OPLS} - \underline{B})$

$\xrightarrow{D} N_p(\underline{0}, D \Sigma_y D^T)$ where by HW 7 #1:

$$D = \begin{pmatrix} \frac{\partial \lambda \mu_1}{\partial \lambda} & \frac{\partial \lambda \mu_1}{\partial \mu_1} & \cdots & \frac{\partial \lambda \mu_1}{\partial \mu_p} \\ \vdots & & & \\ \frac{\partial \lambda \mu_p}{\partial \lambda} & \frac{\partial \lambda \mu_p}{\partial \mu_1} & \cdots & \frac{\partial \lambda \mu_p}{\partial \mu_p} \end{pmatrix} = (\underline{\mu} \;\; \lambda I_p).$$

c) Let A be a $K \times P$ constant matrix with full rank K: $(1 \le K \le P)$ where

$$A\underline{B} = A\underline{m} = \underline{0}.$$

Then $\sqrt{n}\left(A\hat{\underline{B}}_{OPLS} - \underline{0}\right) \xrightarrow{D} N_K\left(\underline{0}, \lambda^2 A \Sigma_{\underline{m}} A^T\right)$

see HW 1b).

d) Estimator: $\hat{\lambda}^2 A \hat{\Sigma}_{\underline{m}} A^T$

where $\hat{\Sigma}_{\underline{m}} = \hat{\Sigma}_{\underline{w}} = \hat{\Sigma}_{\underline{z}} = \frac{1}{n-1}\sum_{i=1}^{n} \underbrace{x_i (y_i - \bar{y})}_{\underline{z}_i}$

e) $H_0: A\underline{B} = \underline{0}$ iff $H_0: A\underline{m} = \underline{0}$.

$H_0: A\underline{B} = \underline{0}$: slightly harder test if $n >> p$

$H_0: A\underline{m} = \underline{0}$: easier test

20) For high dimensional data and big data, often $\frac{n}{p}$ is not large, eg $p >> n$.

with $\hat{\underline{M}} = \widehat{cov}(\underline{x}, Y)$, can still do a lot of testing because

$\sqrt{n}(\hat{\underline{M}} - \underline{M}) \xrightarrow{D} N_p(\underline{0}, \underline{\Sigma}_W)$ and $\underline{\Sigma}_W$ is not an inverse matrix

(MLEs and OLS estimate $\underline{\Sigma}$ with an inverse matrix like $(\underline{X}^T\underline{X})^{-1}$ or $I_1^{-1}(\hat{\beta})$.)

If $n < 5p$ we can't get a good estimator of $cov(\underline{x}) = cov(\underline{\Sigma}_{XY}) = \underline{\Sigma}_W$, but we can get good nonsingular estimators of $cov(\underline{\Sigma}_{UY}) = cov\{(\hat{m}_{i1}, \ldots, \hat{m}_{ik})^T\}$ with $\underline{U}_i = (x_{i1}, \ldots, x_{ik})^T$ where $n \geq 10k$: use the sample covariance matrix of the vectors $\underline{U}_i(Y_i - \bar{Y})$. Hence we can test hypotheses like $H_0: \beta_i = 0$ or $H_0: \beta_i - \beta_j = 0$.

21) Variable selection is useful

in low and high dimensions.

Let $Y \perp\!\!\!\perp \underline{X} \mid \underline{\beta}^T \underline{X}$ eg $Y = \underline{X}^T \underline{\beta} + \varepsilon$

or $Y \sim \text{Pois}[\exp(\underline{X}^T \underline{\beta})]$. Assume a constant β_1, $x_{i1} \equiv 1$ is always in the model.

A model for variable selection is

$$\underline{X}^T \underline{\beta} = \underline{X}_S^T \underline{\beta}_S + \underline{X}_E^T \underline{\beta}_E = \underline{X}_S^T \underline{\beta}_S \qquad (*)$$

where $\underline{X} = (\underline{X}_S^T, \underline{X}_E^T)^T$, $\underline{X}_S$ is $a_S \times 1$

$\underline{X}_E$ is $(p - a_S) \times 1$. Given $\underline{X}_S$ is in the model, $\underline{\beta}_E = \underline{0}$ and E denotes the subset of terms in the model that can be eliminated given S is in the model. Often take S to be the unique subset of important variables that should be in the model.

Since S is unknown, candidate subsets are examined. Let $\underline{x}_I$ be a candidate $a \times 1$ vector indexed by I and let $\underline{x}_O$ be the vector of predictors out of the model.

Then $\underline{x}^T\underline{\beta} = \underline{x}_I^T\underline{\beta}_I + \underline{x}_O^T\underline{\beta}_O$.

If $S \subseteq I$ and (*) holds, then

$$\underline{x}^T\underline{\beta} = \underline{x}_S^T\underline{\beta} = \underline{x}_S^T\underline{\beta}_S + \underline{x}_{I/S}^T\underline{\beta}_{I/S} + \underline{x}_O^T\underline{0}$$

$$= \underline{x}_I^T\underline{\beta}$$

where $\underline{x}_{I/S}$ denotes the predictors in I that are not in S. (slash)

I is a submodel $(Y \perp\!\!\!\perp \underline{x}_I \mid \underline{x}_I^T\underline{\beta})$, and the full model is a submodel.

22) Forward selection I_1 uses $\underline{x}_1^* \equiv 1 = \underline{x}_1$ the constant, I_2 uses $\underline{x}_1^*, \underline{x}_2^*$ where $\underline{x}_2^*$ minimizes criterion $C(\underline{x}_1^*, \underline{x}_j^*)$, $j \in \{2, \ldots, p\}$.

I_K uses $\underbrace{\underline{x}_1^*, \ldots \underline{x}_{K-1}^*}_{\text{in model}}, \underline{x}_K^*$ where

$\underline{x}_K^*$ minimizes $C(\underline{x}_1^*, \ldots \underline{x}_{K-1}^*, \underline{x}_j^*)$

$j \in \{1, \ldots p\} / \{1^*, \ldots K-1^*\}$. (often C = AIC or BIC.)

Forward selection has r models

$I_1, \ldots, I_r$. Often $r = p$ if $n \geq 10p$.

ex) $p = 4$ $x_1 \equiv 1$ corresponding to β_1 is always in the model. Suppose $S = \{1, 2\} = I_2$. There are $J = 2^{p-1} = 8$ possible subsets of $\{1, \ldots p\}$ that always contain 1! $\frac{1}{1}$ integral, $\frac{2}{2}$ integral, ... $\frac{2}{p}$, There are 2^{p-2} subsets that contain S.

$\underline{\hat{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ is obtained by regressing Y on $\underline{x}_{I_7} = (x_1, x_3, x_4)^T$.

Let Imin correspond to the predictors

Selected by the variable selection method such as forward selection or Lasso (later). If $\hat{\underline{\beta}}_I$ is a $\times 1$, form the $p \times 1$ vector $\hat{\underline{\beta}}_{I,0}$ by adding 0's corresponding to the omitted variables.

eg $p=4$ $\hat{\underline{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$. Then

$$\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T.$$

As a statistic, $\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_k,0}$ with probs $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, \ldots, J$

eg $J = 2^{p-1}$.

23] A random vector $\underline{U}$ has a mixture distribution of random vectors $\underline{U}_j$ with probabilities π_j if $\underline{U}$ equals the randomly selected vector $\underline{U}_j$ with probs π_j for $j = 1, \ldots, J$. (The selection process must not change the dist of the $\underline{U}_j$.)

$F_{\underline{U}}(\underline{t}) = \sum_{j=1}^{J} \pi_j F_{\underline{U}_j}(\underline{t})$ where

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

$F_{\underline{U}_j}(\underline{t})$ is the cdf of $\underline{U}_j$.

Suppose both $E(h(\underline{U}))$ and $E(h(\underline{U}_j))$ exist. Then $E[h(\underline{U})] = \sum_{j=1}^{J} \pi_j E[h(\underline{U}_j)]$

$$E[\underline{U}] = \sum_{j=1}^{J} \pi_j E(\underline{U}_j)$$

$$\text{Cov}(\underline{U}) = \sum_{j=1}^{J} \pi_j \text{Cov}(\underline{U}_j) + \sum_{j=1}^{J} \pi_j E[\underline{U}_j][E(\underline{U}_j)]^T - E(\underline{U})[E(\underline{U})]^T.$$

If $E(\underline{U}_j) = \underline{0}$ for $j = 1, \ldots, J$, then

$E(\underline{U}) = \underline{0}$ and $\text{Cov}(\underline{U}) = \sum_{j=1}^{J} \pi_j \text{Cov}(\underline{U}_j)$.

24) 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

$$P(B) = \sum_{k=1}^{J} P(B \cap A_k) = \sum_{k=1}^{J} P(B|A_k) P(A_k)$$

Variant: if $P(A_k) = 0$ define

$P(B|A_k) P(A_k) = 0$.

25) Let $w = k$ if $\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_{k0}}$ where

$P(w = k) = \pi_{kn}$ for $k = 1, \ldots, J$.

Let $\hat{\underline{\beta}}^{c}_{I_{k0}}$ be a random vector from the conditional dist $\hat{\underline{\beta}}_{I_{k0}} | w = k$.

Let $\underline{w}_n = \sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta})$ and

$$\underline{w}_{kn} = \sqrt{n}(\hat{\underline{\beta}}_{I_{k0}} - \underline{\beta}) \,|\, w = k \stackrel{D}{=} \sqrt{n}(\hat{\underline{\beta}}^{c}_{I_{k0}} - \underline{\beta}).$$

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

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 $\hat{\underline{\beta}}^{c}_{I_{k0}}$ with probs π_{kn} and

$\underline{W}_n$ has a mixture dist of the $\underline{W}_{kn}$ with probs π_{kn}.

Proof} $F_{\underline{W}_n}(\underline{t}) = P\{\sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta}) \le \underline{t}\} =$ (B; law of tot prob)

$$\sum_{k=1}^{J} P\left(\sqrt{n}(\hat{\underline{\beta}}_{VS} - \underline{\beta}) \le \underline{t} \mid \underbrace{\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_{k0}}}_{A_k}\right) P\left(\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_{k0}}\right)$$

$$= \sum_{k=1}^{J} P\left\{\sqrt{n}(\hat{\underline{\beta}}_{I_{k0}} - \underline{\beta}) \le \underline{t} \mid \hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_{k0}}\right\} \pi_{kn}$$

$$= \sum_{k=1}^{J} P\left\{\sqrt{n}(\hat{\underline{\beta}}_{I_{k0}^c} - \underline{\beta}) \le \underline{t}\right\} \pi_{kn}$$

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

26} Assume that if $S \subseteq I_j$ where $\dim(I_j) = q_j$, then $\sqrt{n}(\hat{\underline{\beta}}_{I_j} - \underline{\beta}_{I_j}) \xrightarrow{D} N_{q_j}(\underline{0}, V_j)$.

Then $\sqrt{n}(\hat{\underline{\beta}}_{I_{j0}} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, V_{j0})$

where V_{j0} adds rows and columns

of zeroes corresponding to the x_j not in I_j. V_{j0} is singular unless I_j is the full model.

Let $\hat{\beta}_{MIX}$ have a mixture dist of the $\hat{\beta}_{I_{K0}}$ with prob's π_{Kn} but the I_K are randomly selected.

eg is Simulation: generate a data set perform VS and pick I_{K0} and $\hat{\beta}_{VS}$, generate another data set and use I_{K0} to get $\hat{\beta}_{MIX}$.

27] Thi Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$.

Let $\hat{\beta}_{MIX} = \hat{\beta}_{I_{K0}}$ with probs π_{Kn} where $\pi_{Kn} \to \pi_K$ as $n \to \infty$. Denote the positive π_K by π_j. Assume

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

a) $\underset{\sim}{U}_n = \sqrt{n}\left(\hat{\underset{\sim}{B}}_{MIX} - \underset{\sim}{B}\right) \xrightarrow{D} \underset{\sim}{U}$ where

the cdf $F_{\underset{\sim}{U}}(t) = \underbrace{\sum_j \pi_j F_{\underset{\sim}{U}_j}(t)}$.

mixture dist of the $\underset{\sim}{U}_j$

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

$\underset{\sim}{V}_n = A\underset{\sim}{U}_n = \sqrt{n}\left(A\hat{\underset{\sim}{B}}_{MIX} - A\underset{\sim}{B}\right)$

$\xrightarrow{D}$ $A\underset{\sim}{U} = \underset{\sim}{V}$ where $\underset{\sim}{V}$ has

a mixture dist of the $\underset{\sim}{V}_j =$

$A\underset{\sim}{U}_j \sim N_g(\underset{\sim}{0}, A V_{j0} A^T)$.

c) $\hat{\underset{\sim}{B}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\underset{\sim}{B}$; $\sqrt{n}\left(\hat{\underset{\sim}{B}}_{VS} - \underset{\sim}{B}\right) = O_p(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}\left(\hat{\underset{\sim}{B}}_{SEL} - \underset{\sim}{B}\right) \xrightarrow{D} N_p(\underset{\sim}{0}, V_{d0})$

where SEL is MIX or VS.

e) Let $\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_{K0}}$ with probs π_{Kn}.

Assume $\underline{w}_{jn} = \sqrt{n}\left(\hat{\underline{\beta}}^{C}_{I_{j0}} - \underline{\beta}\right) \xrightarrow{D} \underline{w}_j$.

Then $\underline{w}_n = \sqrt{n}\left(\hat{\underline{\beta}}_{VS} - \underline{\beta}\right) \xrightarrow{D} \underline{w}$

where $F_{\underline{w}}(t) = \sum_{j=1}^{J} \pi_j F_{\underline{w}_j}(t)$.

Thus $\underline{w}$ is a mixture dist of the $\underline{w}_j$ with probs π_j.

Proof a) Since $\underline{v}_n$ has a mixture dist of the $\underline{v}_{Kn}$ with probs π_{Kn}, the cdf of $\underline{v}_n$ is $\sum_K \pi_{Kn} F_{\underline{v}_{Kn}}(t) \rightarrow$

$F_{\underline{v}}(t) = \sum_j \pi_j F_{\underline{v}_j}(t)$ at continuity points of $F_{\underline{v}_j}(t)$ as $n \rightarrow \infty$.

b) Since $\underset{\sim}{U}_n \xrightarrow{p} \underset{\sim}{U}$, $A\underset{\sim}{U}_n \xrightarrow{p} A\underset{\sim}{U}$.

c) Selecting from a finite # of $\sqrt{n}$ consistent estimators (even on a set that goes to 1 in prob), results in a $\sqrt{n}$ consistent estimator by prob. See HW7 #4.

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

e) proof almost the same as a) □

28) The assumption $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ is important. The assumption has been proved for AIC and BIC for MLR, GLMs, AR(P) time series. A necessary condition for estimators (like lasso and adaptive

net to be consistent is that $P(S \subseteq I_{min}) \to 1$, so OLS or PLS after elastic net or lasso satisfies the assumption for MLR under mild conditions,

29] Let $\underset{\sim}{x} = (1, \underset{\sim}{u}^T)^T$ and $\underset{\sim}{Y} = \underset{\sim}{x}\underset{\sim}{\beta} + \underset{\sim}{e}$, MLR. Let W be the matrix of standardized $\underset{\sim}{u}$ so that

$R_{\underset{\sim}{u}} = \frac{W^T W}{n}$ is the sample correlation matrix of the $\underset{\sim}{u}$. Let $\underset{\sim}{Z} = \underset{\sim}{Y} - \bar{Y}$. Then regression through the origin is used for

$\underset{\sim}{Z} = W \underset{\sim}{\eta} + \underset{\sim}{e}$.

$\hat{\underset{\sim}{Y}} = \bar{Y} + \hat{\underset{\sim}{Z}}$. W does not contain a column of 1s. $\hat{\underset{\sim}{\beta}}$ can be obtained from $\hat{\underset{\sim}{\eta}}$ and $\bar{Y}$.

This method is used so that people who use the same units of measurement (eg could use feet or meters etc), get the same answer from the software.

3C) Assume $R_{\underline{U}} = \frac{W^T W}{n} \xrightarrow{P} V^{-1}$.

Then $V^{-1} = \rho_{\underline{U}}$ = pop corr matrix of the nontrivial predictors $\underline{U}_i$ if the $\underline{U}_i$ are iid.

Let $H = W (W^T W)^{-1} W^T = (h_{ij})$.

OLS CLT: Assume $\max_{i=1,\ldots,n} h_{ii} \xrightarrow{P} 0$

as $n \to \infty$. Then

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

(↑ no longer $\hat{\Sigma}_{xy}$)

3D) $\underline{Y} = X\underline{\beta} + \underline{e}$, $\underline{Z} = W\underline{\eta} + \underline{e}$