24) In high dimensions, usually can't compute $\hat{\beta}_S$ & β_{FULL} and S may not be a subset of I_j for any $j = 1, \ldots, J$.

25) Data splitting assumes the data follows a regression model and that the cases are ind. Randomly divide the data into two sets: the training set or modeling set H has n_H cases and the validation set V has the remaining $n_V = n - n_H$ cases $i_1, \ldots, i_{n_V}$. Use set H to build a model I_H with a predictors. For set H, looking at the data is allowed. Variable selection can be

used. (On a full data set on n cases, looking at the response invalidates hypothesis testing, and variable selection inference is difficult.)

Then fit model I_H with a predictors with the validation set. If $n_v \geq Ja$ ($J \geq 5, 10$ etc), do standard model checks and inference.

26] Drawbacks: i) n_v cases instead of n are used, so loss of efficiency compared to choosing model I_H without looking at the response.

ii) Models that are much better than I_H may exist, especially in high dimensions

27] Data splitting is easy to do with high dimensional data, and iid cases are not needed.

28] * The R sample function can be used to get a permutation of $1, \ldots, n$. Suppose $a_1, \ldots, a_n$ represent the cases.

$n \in 10$

sample(1:n, n) and sample(1:n) work.

SEE HW5

($n_H \simeq \frac{n}{2}$ is common)

ex] $n = 10$ $a_1, \ldots, a_{10}$, $n_H = 5$

sample(1:10)

4 1 9 6 10 | 3 5 2 8 7

H V

$a_4, a_1, a_9, a_6, a_{10}$ in H

could use sample(1:10, 5) (n_H)

1 7 5 9 8

a_1, a_7, a_5, a_9, a_8 in H.

29] I like to use student names on quizzes and exams.

30] OLS MLR variable selection $n >> p$ $\underset{\sim}{Y} = \underset{\sim}{X}^T \underset{\sim}{\beta} + \underset{\sim}{e}$

Let $MSE(I) = \frac{SSE(I)}{n-k} = \frac{\sum (y_i - \hat{y}_{i(I)})^2}{n-k}$ where

$\underset{\sim}{\beta}_I$ contains constant is $k \times 1$.

Then $Cp(I) = \frac{SSE(I)}{MSE} + 2k - n$

$= (p-k)(F_I - 1) + k$ where F_I is the statistic for testing $H_0: \underset{\sim}{B}_0 = \underset{\sim}{0}$ out of the model in the model $Y = \underset{\sim}{X}_I^T \underset{\sim}{B}_I + \underset{\sim}{X}_0^T \underset{\sim}{B}_0 + e$.

Also $MSE = MSE(F) = \hat{\sigma}^2 =$ full model MSE.

31] Let $\underset{\sim}{r}$ be the full model residuals and $\underset{\sim}{r}_I$ be the submodel residuals:

$r_i(I) = \hat{Y}_i(I)$

$\underset{\sim}{Y} = \underset{\sim}{X}_I \underset{\sim}{B}_I + \underset{\sim}{e}$, $r_{iI} = Y_i - \hat{Y}_{iI}$, $i = 1, \ldots, n$.

suppose every submodel I contains a constant and $\underset{\sim}{X}$ is a full rank matrix. For OLS,

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}$$

(p = full model)

$$F_I = \left[\frac{SSE(I) - SSE}{(n-k) - (n-p)}\right] \Big/ \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right].$$

32] Consider all submodels I (considered by the US method) with $\underset{\sim}{B}_I$ $k \times 1$.

The model that minimizes $C_p(I)$ maximizes $\text{corr}(r, r_I)$.

and MSE(I)

33] Of the J models considered, eg from forward selection, let I_{min} minimize $C_p(I)$.

34] When $H_0: \underset{\sim}{\beta}_0 = \underset{\sim}{0}$ is true,

$$(p-k) F_I \xrightarrow{D} \chi^2_{p-k} \quad \text{and}$$

$$(p-k)(F_I - 1) + k = C_p(I) \xrightarrow{D} \chi^2_{p-k} + 2k - p$$

if the errors are iid with constant variance σ^2.

forward sel and backward elim

35] Suppose the full model F is one of the models considered so $S \subseteq F$. Let $\underset{\sim}{Y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{e}$, $\underset{\sim}{\beta} \leftarrow \underset{\sim}{\beta}_F$.

$C_p(I_{min}) \leq C_p(F) = p$ and $corr(\underset{\sim}{r}, \underset{\sim}{r}_{I_{min}}) \to 1$ as $n \to \infty$. HW5

If $\underset{\sim}{x}^T\underset{\sim}{\beta} = \underset{\sim}{x}_S^T \underset{\sim}{\beta}_S$ and $P(S \subseteq I_{min})$ does not go to one, then $corr(\underset{\sim}{r}, \underset{\sim}{r}_{I_{min}})$ would not go to one. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ (assuming S is unique and has minimal dimension).

ex] Assume $S = I_i$ is the model that deletes predictor x_i. Then $\underset{\sim}{\beta}_{I_i}$ is $(p-1) \times 1$. Consider all subsets selection. Then $S \subseteq I_i$ and $S \subseteq F$ so only π_S and π_F are positive since $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Since $C_p(F) = p$, $S = I_i$ is selected if $C_p(S) < p$.

By 34) $C_p(S) \xrightarrow{D} \chi^2_{p-k} + 2k - p \underset{\text{if } k=p-1}{=} \chi^2_1 + p - 2.$

Hence $\pi_S = P(\chi^2_1 + p - 2 < p) = P(\chi^2_1 < 2)$

and $\pi_F = 1 - \pi_S$. This result will also hold for backward elimination. For forward selection, expect $\pi_S < P(\chi_1^2 < 2)$ and $\pi_F > 1 - P(\chi_1^2 < 2)$.

(S will often not be 1 of the 3 models for forward sel.)

Data splitting Prediction Region

36] Suppose there is training data $\underset{p\times 1}{\underline{x}_1}, \ldots, \underline{x}_n$ and test data $\underline{x}_f$ (a future value).

A large sample $100(1-\delta)\%$ prediction region is a set A_n such that $P(\underline{x}_f \in A_n)$ is eventually bounded below by $1-\delta$ as $n \to \infty$. (α is often used)

37] If $\underline{x}_f$ has a pdf, we often want $P(\underline{x}_f \in A_n) \to 1-\delta$ as $n \to \infty$. A prediction interval $[L_n, U_n] = A_n$ is a prediction region where $p = 1$ ($x_f = X_f$).

38] If $\underline{x}_1, \ldots, \underline{x}_n, \underline{x}_f$ are iid $N_p(\underline{\mu}, \Sigma)$, Σ^{-1} exists, then the large sample $100(1-\delta)\%$ classical prediction region is $\{\underline{z} : D^2_{\underline{z}}(\bar{\underline{x}}, S) \le \chi^2_{p,1-\delta}\}$.

$\hat{\theta} = \hat{\Sigma}_{\underline{x}}$.

39] Let $q_n = \min(1-\delta+0.05, 1-\delta+\frac{p}{n})$, $\delta > 0.1$

$q_n = \min(1-\frac{\delta}{2}, 1-\delta+10\frac{\delta p}{n})$, $\delta \leq 0.1$.

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$, set $q_n = 100(1-\delta)$. Let $D_{(U_n)}$ be the $100\,q_n$ th sample quantile of the $D_i = D_{\underline{x}_i}$

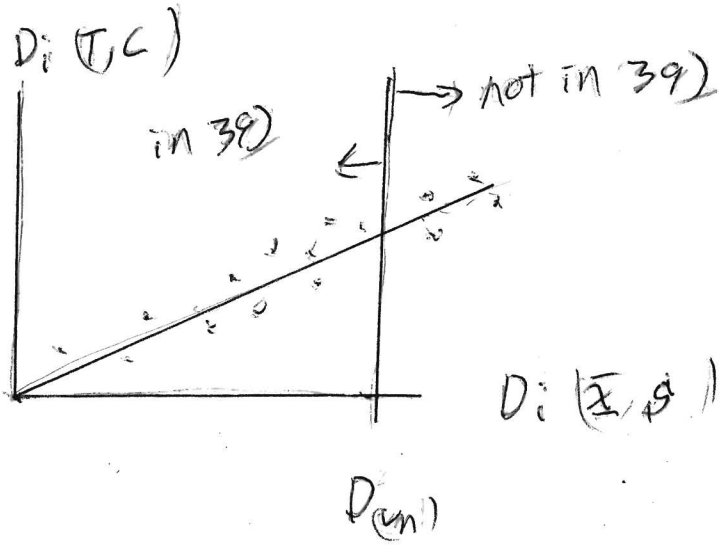
where $D^2_{\underline{x}_i} = (\underline{x}_i - \bar{\underline{x}})^T S^{-1} (\underline{x}_i - \bar{\underline{x}})$, $i = 1, \ldots, n$.

If $\underline{x}_1, \ldots, \underline{x}_n, \underline{x}_f$ are iid and $\Sigma_{\underline{x}}^{-1}$ exists, then the large sample $100(1-\delta)\%$ nonparametric prediction region for $\underline{x}_f$ is

$$\{\underline{z} : D^2_{\underline{z}}(\bar{\underline{x}}, S) \leq D^2_{(U_n)}\}.$$

40] 39) uses $D^2_{(U_n)}$ while 38) uses $\chi^2_{p,1-\delta}$.

39) contains $\approx 100\,q_n\%$ of the training data.



If $\delta = 0.95$, $n \leq 20p$, then $q_n = 0.975$.
Methods usually fit training data better than test data.

41] If X_1 and $X_2 = X_f$ are iid random variables, then $P(X_1 \le X_2) = P(X_2 \le X_1) \ge 0.5$.

So $[X_1, \infty)$ and $(-\infty, X_1]$ are 50% prediction intervals for $X_2 = X_f$.

42] Use data splitting to divide the iid training data $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ into 2 sets H and V. Compute (T_H, C_H) from the n_H cases in H where C_H^{-1} exists eg $C_H = I_p$. Then compute the squared validation distances $D_j^2 = D_{\underset{\sim}{x}_{ij}}^2 = (\underset{\sim}{x}_{ij} - T_H)^T C_H^{-1} (\underset{\sim}{x}_{ij} - T_H)$

for the cases $\underset{\sim}{x}_{i_1}, \ldots, \underset{\sim}{x}_{i_{n_V}}$ in V. Let $(D^2_{(U_V)}$ be the U_Vth order statistic of the D_j^2 where $U_V = \min(n_V, \lceil (n_V+1)(1-\delta) \rceil)$ where the ceiling function $\lceil x \rceil$ rounds x up to the nearest integer: $\lceil 7.7 \rceil = \lceil 8 \rceil = 8$.

Then the large sample $100(1-\delta)\%$ data splitting prediction region for $\underset{\sim}{x}_f$ is $\{\underset{\sim}{z}: D^2_{\underset{\sim}{z}}(T_H, C_H) \leq D^2_{(U_V)}\}$.

433 42) can be used for high dimensional data, if $\hat{\underset{\sim}{\Sigma}}_x^{-1}$ does not exist, and for estimators with little theory such as $(T_H, C_H) = (T_{COVMB2}, C_{COVMB2})$. The "actual coverage $\approx \frac{U_V}{n_V+1} \geq \frac{\lceil (n_V+1)(1-\delta) \rceil}{n_V+1}$

n_V	δ	$U_V = \min(n_V, \lceil (n_V+1)(1-\delta) \rceil)$	$\frac{U_V}{n_V+1}$
1	1/2	$1 = \lceil 2 \frac{1}{2} \rceil$	$\frac{1}{2} = 0.5$
2	1/3	$2 = \lceil 3 \frac{2}{3} \rceil$	2/3
3	1/4	$3 = \lceil 4 \frac{3}{4} \rceil$	3/4
⋮			
9	1/10	$9 = \lceil 10 \cdot 9 \rceil$	9/10
10	1/10	$10 = \min(10, \lceil 11 \frac{9}{10} \rceil)$	$10/11 \approx 0.9091$
19	1/20	$19 = \lceil 20 \cdot \frac{19}{20} \rceil$	19/20
20	1/20	$20 = \min(20, \lceil 21 \cdot \frac{19}{20} \rceil)$	$20/21 \approx 0.9524$
99	1/20	$95 = \lceil 100 \frac{19}{20} \rceil$	95/100
100	1/20	$96 = \lceil 101 \frac{19}{20} \rceil$	$96/101 \approx .9505$
999	1/20	$950 = \lceil 1000 \cdot 19/20 \rceil$	950/1000
1000	1/20	$951 = \lceil 1001 \cdot 19/20 \rceil$	$951/1000 = .951$

n_V	α	UV	cov	
600	$\frac{1}{20}$	571	571/601	.950

If $n_V \geq 19$, can get "coverage" close to 0.95.

ex) $\underline{x}_1 = \cdots = \underline{x}_n = \underline{0} = \underline{x}_f = T$, C nonsingular

coverage = 100%

$D_i^2 = 0$ for $i = 1, \ldots, n$

Closed region

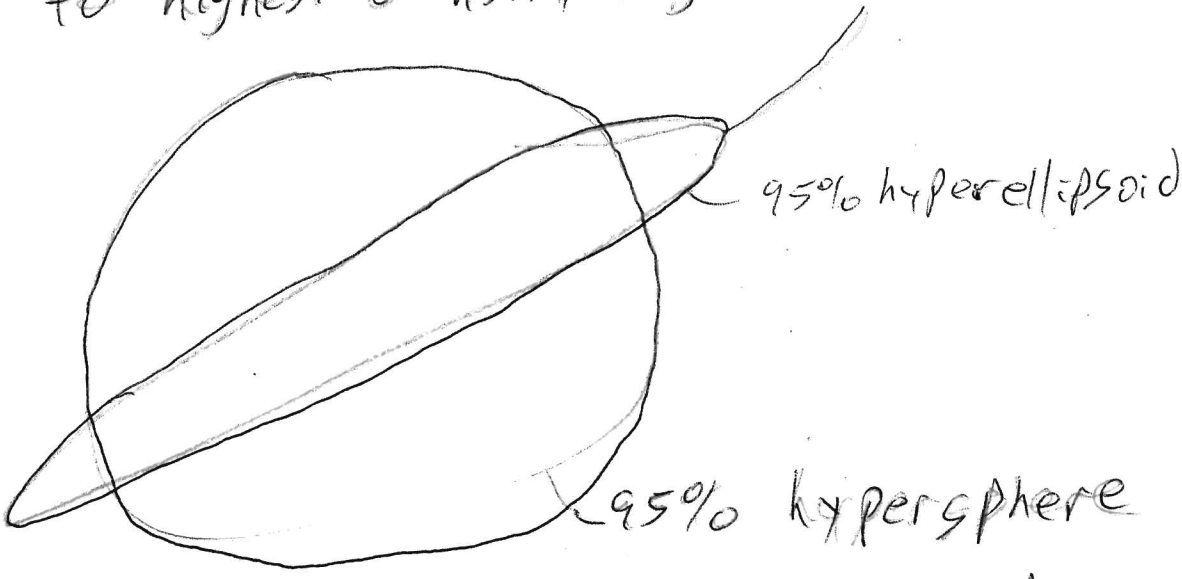
$D^2_{(v,1)} = 0$

pred region $= \{\underline{z} : \underline{z} = \underline{0}\}$.

44) May need $n_H \geq 20p$ to get a good estimate of C_H (if $C_H \neq I_p$).

May need $n_V \geq 50$ for $D_{(UV)}$ to be a good estimator of the "percentile" $D_{(UV)} \approx D_{(n(1-\alpha))}$.

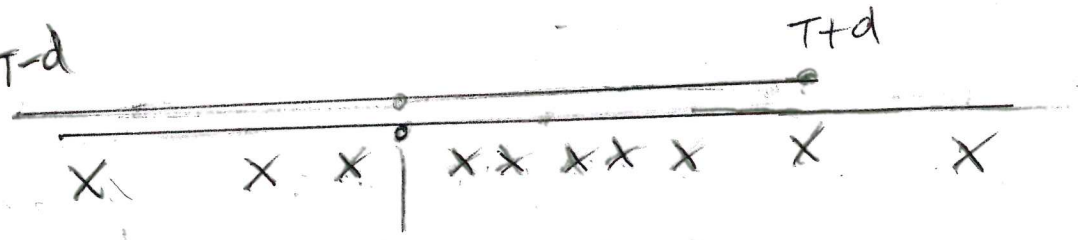
Drawback: volume of 42) may be huge compared to "highest density region volume."



45) If $p = 1$, the 42) prediction interval

is the closed interval $[T-d, T+d]$

where $d \geq 0$ is the smallest number such that the PI contains $\geq U_V$ cases.



$T = \text{MED}(n_H)$ $\quad n_V = 10, U_V = 9$

(cases shown from V, not H)

46] Proof that 42] is a prediction region.

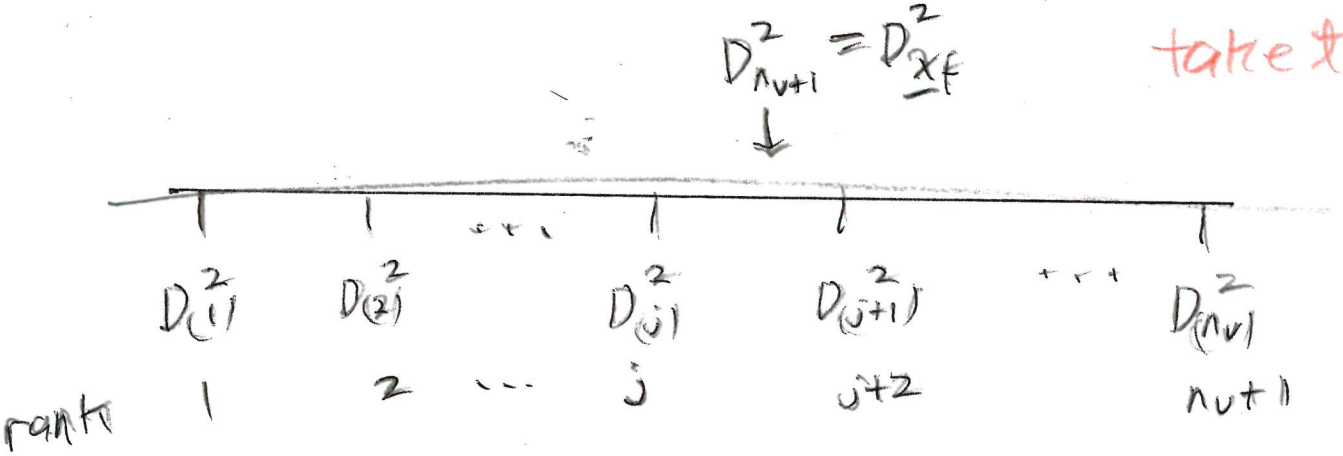
Assume $\underline{x}_1, \ldots, \underline{x}_n, \underline{x}_f$ are iid where $\underline{x}_f = \underline{x}_{n+1}$. Compute (T, C_H) from the cases in H. Consider the squared validation distances D_j^2 for $j = 1, \ldots, n_V$ and $D_{n_V+1}^2$ for case $\underline{x}_f$ (unobserved). Since the $n_V + 1$ cases are iid, the prob that D_t^2 has rank K for $K = 1, \ldots, n_V+1$ is $\frac{1}{n_V+1}$ for each t (i.e. the ranks follow a discrete Uniform dist).

$t = 1, \ldots, n_V+1$

Let $D^2_{(i)}$ be the order statistics without using the unknown squared "validation distance" $D^2_{n_v+1} = D^2_{\underline{x}_f}$. Then

$D^2_{(i)}$ has rank i if $D^2_{(i)} < D^2_{n_v+1}$

but rank $i+1$ if $D^2_{(i)} > D^2_{n_v+1}$.



Thus $D^2_{(U_v)}$ has rank U_v+1 if $D^2_{\underline{x}_f} < D^2_{(U_v)}$,

and $P\left(\underline{x}_f \in \left\{\underline{z} : D^2_{\underline{z}}(T_H, C_H) \le D^2_{(U_v)}\right\}\right)$

$= P\left(D^2_{\underline{x}_f} \le D^2_{(U_v)}\right) \ge \frac{U_v}{n_v+1} \to 1-\delta$ as $n_v \to \infty$.

If there are no tied ranks,

$P\left(D^2_{\underline{x}_f} \le D^2_{(U_v)}\right) = P\left(D^2_{\underline{x}_f} < D^2_{(U_v)}\right) =$