58] Let the Sufficient predictor
$E(Y|\underset{\sim}{x}_I) = SP = \underset{\sim}{x}_I^T \underset{\sim}{\beta}_I$.

So $E(Y|\underset{\sim}{x}) = SP = \underset{\sim}{x}^T \underset{\sim}{\beta}$ for the full model.

Let the estimated sufficient predictor
$ESP = \hat{Y} = \underset{\sim}{x}^T \hat{\underset{\sim}{\beta}}$.
$ESP(I) = \hat{Y}_I = \underset{\sim}{x}_I^T \hat{\underset{\sim}{\beta}}_I$.

59] For $\underset{\sim}{x}_S$ and $\underset{\sim}{x}_I$, suppose $S \subseteq I$
and that 42] holds: $Y = \underset{\sim}{x}^T \underset{\sim}{\beta} + e = \underset{\sim}{x}_S^T \underset{\sim}{\beta}_S + e$

Then $SP = \underset{\sim}{x}^T \underset{\sim}{\beta} = \underset{\sim}{x}_S^T \underset{\sim}{\beta}_S = \underset{\sim}{x}_S^T \underset{\sim}{\beta}_S + \underset{\sim}{x}_{I/S}^T \underset{\sim}{\beta}_{I/S} + \underset{\sim}{x}_O^T \underset{\sim}{\beta}_O$
$= \underset{\sim}{x}_I^T \underset{\sim}{\beta}_I$, and $corr(\underset{\sim}{x}^T \underset{\sim}{\beta}, \underset{\sim}{x}_I^T \underset{\sim}{\beta}_I) = 1$.
Here $\underset{\sim}{x}_{I/S}$ is the vector of predictors in $\underset{\sim}{x}_I$ but not in $\underset{\sim}{x}_S$.
So submodel I is worth considering if the full model $Y = \underset{\sim}{x}^T \underset{\sim}{\beta} + e$ is good and $corr(\underset{\sim}{x}^T \hat{\underset{\sim}{\beta}}, \underset{\sim}{x}_I^T \hat{\underset{\sim}{\beta}}_I) \geq 0.95$.

60] full model: $\hat{Y}$, $r = Y - \hat{Y}$, $\hat{\underset{\sim}{\beta}}$
submodel I: $\hat{Y}_I$, $r_I = Y - \hat{Y}_I$, $\hat{\underset{\sim}{\beta}}_I$
$\hat{Y}_{I,i} = \hat{\underset{\sim}{\beta}}_I^T \underset{\sim}{x}_{Ii}$, $r_{Ii} = Y_i - \hat{\underset{\sim}{\beta}}_I^T \underset{\sim}{x}_{I,i}$

61] * p100 If you have one data set, build a full model for that data set, then

perform variable selection to obtain a final (sub)model, then an extreme amount of data snooping was used. The final model may be useful for exploratory data analysis and description, but the pvalues for the t tests and Anova F test are likely too small while the pval for the partial F test that uses the final model as the reduced model is likely too high. The final model tends to fit the data set from which it was built better than future obs's. PIs tend to be too short so coverage < nominal coverage eg 95%.

62] Inference such as tests and PI's are valid if the final MLR model is selected before gathering data, not after using variable selection to build a final model. Data splitting is useful.

63] p101 Leaving out important predictors is called underfitting. Such models tend to be very poor and may violate linearity ($Y = \underline{x}_I^T \underline{\beta}_I + e$) and constant variance ($V(Y_i | \underline{x}_I) \equiv \sigma^2$) assumptions.

64] After performing variable selection, the tests and CIs are useful for exploratory

purposes and description, but are not valid for inference.

65] p101 Having unnecessary predictors in the model I is called overfitting or fitting noise.

Let 4.2) hold with $S \subseteq I$.

If $\underset{p\times 1}{\underline{Y}} = \underset{p\times 1}{\underline{x}^T\underline{\beta}} + e = \underset{1\times j}{\underline{x}_I^T}\underset{j\times 1}{\underline{\beta}_I} + e = \underset{1\times k}{\underline{x}_S^T}\underset{k\times 1}{\underline{\beta}_S} + e$

$$\frac{1}{n}\sum V(\hat{Y}_{Ii}) = \frac{\sigma^2 j}{n} > \frac{\sigma^2 k}{n} = \frac{1}{n}\sum V(\hat{Y}_{Si}).$$

↑ worst when $j=p$ (consider $n \approx p >> k$)

66] p109 *Rules of thumb 3.7 Assume full model has good response and residual plots and that $n > 5p$. Let subset I have k predictors including a constant. Let I_{min} be the min Cp model and let I_I be the model with the fewest predictors satisfying $C_p(I_I) \leq C_p(I_{min}) + 1$.

* Do not use more predictors than I_I.

Often can't have the following 10 rules hold simultaneously: submodel I is good if

i) the response and residual plots for the submodel look like the response and residual plots for the full model

ii) $\text{corr}(\hat{Y}, \hat{Y}_I) \geq 0.95$

iii) The plotted points in the FF plot cluster tightly about the identity line

35.5

iv) pval ≥ 0.01 (not 0.05) for the partial F test that uses I as the reduced model.

v) want $K \leq n/10$

vi) the plotted points in the RR plot should cluster tightly about the identity line.

vii) want $R^2(I)$ close to $R^2(full)$ ($R^2(I) \leq R^2(full)$ since adding predictors does not decrease R^2)

viii) want $Cp(I_{min}) \leq Cp(I) \leq \min(2k, p)$ with no big jumps in Cp (the increase should be less than 4) as variables (with 1df) are deleted.

ix) want hardly any predictors with pval > 0.05

x) want few predictors with $.01 \leq$ pval $\leq .05$

Compa E2 problem

ex problem 3.3	L1	L2	L3	L4
# predictors	10	5	4	3
# with $.01 \leq$ pval $\leq .05$	0	1	0	0
# with pval $> .05$	8	0	0	0
$R_I^2 = R^2(I)$	.655	.650	.648	.630
corr $(\hat{Y}, \hat{Y}_I)$	1.0	.996	.992	.981
Cp (I)	10.0	4.0	5.60	13.81
pval for partial F test	1.0	.550	.272	.015

L_1 is the full model, L_2 is the min Cp model. a) which model is I_I? b) which other models, if any, should be looked at?

Soln a) $I_I = L_2$ since $5.6 \leq Cp(L_3) > Cp(I_{min}) + 1 = 5$

b) L_3 since $5.6 < Cp(I_{min}) + 4$ (and $Cp(L_4) > 2k$)

§2.5 and §3.4.1 certain

Stripped notes 36, 37 for now

82) Prediction intervals and Prediction regions applied to a bootstrap sample can be used as confidence intervals and confidence regions.

For HW7

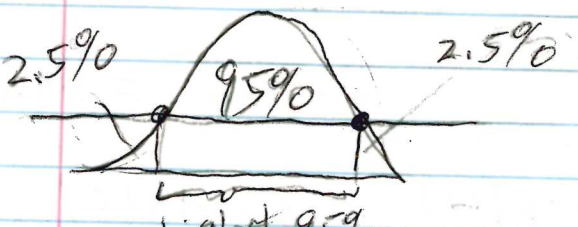
83) Let $Y_1, \ldots, Y_n$ be the data.

The order statistics are

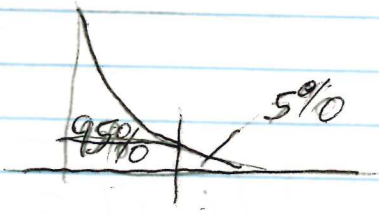
$$Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$$

$Y_{(1)}$: min $Y_{(n)}$: max

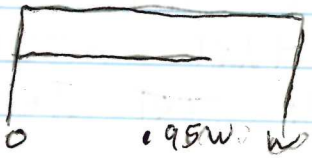
84) The highest $100(1-\delta)\%$ density region of a pdf is found by moving a horizontal line down from the top of the pdf so that the line "intersects the pdf" at one or more intervals and the sum of the areas under the pdf corresponding to the intervals is $1-\delta$



symmetric unimodal lop off $100\frac{\delta}{2}\%$ from both sides

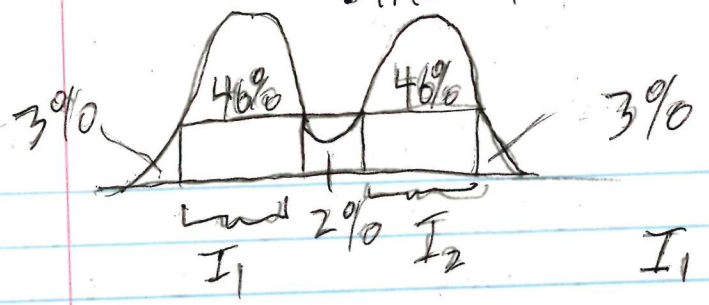


Unimodal right skewed lop off $100\delta\%$ from the right tail



$U(0,w)$ any line within pdf of length $0.95w$ works

bimodal



$I_1 \cup I_2$ = highest 92% region

85] Consider intervals that contain c cases:

$[Y_{(1)}, Y_{(c)}], [Y_{(2)}, Y_{(c+1)}], \ldots, [Y_{(n-c+1)}, Y_{(n)}]$. Compute

$Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, \ldots, Y_{(n)} - Y_{(n-c+1)}$.

Then $\text{Shorth}(c) = [Y_{(s)}, Y_{(s+c-1)}]$ is the closed interval with the shortest length.

ex] Know for E2. Let $c = 4$. Data below has $n = 7$. Find Shorth(4).

see HW 6

0 1 6 9 10 11

$6 = 3 + 6 - 0$

$8 = 9 - 1$

$7 = 10 - 3$

$5 = 11 - 6$ = shortest length

intervals containing $c = 4$ cases

$\boxed{[6, 11] = \text{Shorth}(4) = \text{Shorth}(c).}$

86] If $\frac{c}{n} \to 1 - \delta$, the Shorth(c) interval estimates the highest density $100(1-\delta)\%$ region if that region is an interval. (unimodal pdf that decreases from the mode).