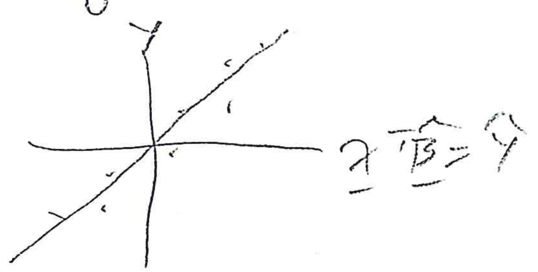
Underfitting is a serious problem:

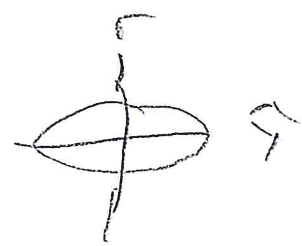
$$Y = x_I^T \beta_I + e_I \qquad \underbrace{var(e_I)}_{\text{may not be constant, could depend on case } i} > var(e) = \sigma^2$$

and the model may no longer be linear. Check with response and residual plots.

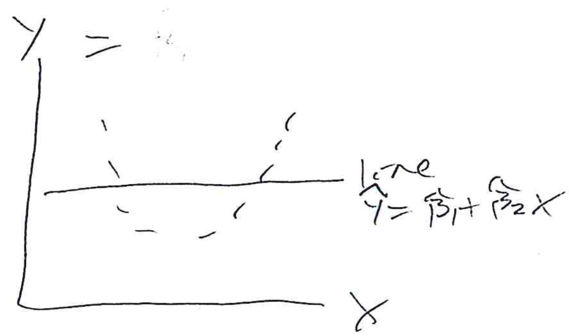
ex] $Y = \beta_1 + \beta_2 x + \beta_3 x^2 + e$

$SP = x^T \beta$ is a hyperplane that is a quadratic in x.





If $SP(I) = \beta_1 + \beta_2 x$ = line in x



expect $\hat{\beta}_2 \approx 0$

8] Forward selection forms a sequence of submodels $I_1, \ldots, I_m$, eg $m = \min(p, \lceil \frac{n}{5} \rceil)$.

I_1 uses $x_1^* = x_1$, a constant but no nontrivial predictors.

To form I_2, consider all models I with 2 predictors including x_1^*. Compute $SSE(I)$ $= RSS(I) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i(I))^2$. Let I_2 minimize

$SSE(I)$, and use x_1^*, x_2^*. In general, to form I_j, consider all models I with j predictors including $x_1^*, \ldots, x_{j-1}^*$. Let I_j minimize $SSE(I)$ and use predictors $x_1^*, \ldots, x_j^*$.

9)

I_j		# models fit to find I_j
I_1	$x_1^* = x_1$	0 (or 1: often do not fit this model)
I_2	x_1^*, x_2^*	$p-1$
I_3	x_1^*, x_2^*, x_3^*	$p-2$
$\vdots$		
I_p	$x_1^*, \ldots, x_p^*$	1

$1 + 2 + \cdots + p-1 = \frac{p(p-1)}{2}$.

10) Need to choose the final model from the sequence of m models $I_1, \ldots, I_m$.

Let [illegible] and [illegible] be [illegible] for a given data set, p, n, and $\underbrace{\hat{\sigma}^2 = MSE}_{\text{full model}}$ act as

constants. A criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion. Let Criterion

$$C_S(I) = SSE(I) + a\, k_n\, \hat{\sigma}^2.$$

$C_p(I) = AIC_S(I)$ uses $k_n = 2$ while $BIC_S(I)$ uses $k_n = \log(n)$. Want $n \geq Jp$, $J \geq 5$, preferably $J \geq 10$. The following criterion still need $\frac{n}{p}$ large.

$$AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a$$

$$BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n),$$

The EBIC criterion may work when $\frac{n}{p}$ is not large.

$$EBIC(I) = BIC(I) + 2 \log\left[\binom{p}{a}\right].$$

$$\log\left[\binom{p}{a}\right] = \underbrace{\log[p!]}_{\text{constant for a given data set.}} - \log[(p-a)!] - \log[a!]$$

11) If $\underline{\hat{\beta}}_{VS} = \underline{\hat{\beta}}_{I_{min},0}$ is a consistent estimator of $\underline{\beta}$, (HB:S) then the probability that I_{min} underfits $\to 0$ as $n \to \infty$. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, a condition that holds for C_p, AIC, BIC lasso variable selection and elastic net variable selection.

12) If $S \subseteq I$ then $\underline{\hat{\beta}}_{I,0}$ is a $\sqrt{n}$ consistent estimator of $\underline{\beta}$ for $Y = \underline{x}^T\underline{\beta} + e = \underline{x}_S^T\underline{\beta}_S + e$. Since there are at most 2^p regression models I that contain S and the prob that I_{min} picks one of these models goes to 1, $\underline{\hat{\beta}}_{VS} = \underline{\hat{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\underline{\beta}$ if $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$.

§11.7 13) A random vector $\underline{u}$ has a mixture distribution of random vectors $\underline{u}_j$ with probabilities π_j if $\underline{u}$ equals $\underline{u}_j$ with probabilities π_j for $j = 1, \ldots, J$ (and the selection mechanism does not change the distribution of the $\underline{u}_j$). Let $\underline{u}$ and $\underline{u}_j$ be random vectors, then the cdf of $\underline{U}$

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

$$E[h(\underline{U})] = \sum_{j=1}^{J} \pi_j \, E[h(\underline{U}_j)] \text{ and } E(\underline{U}) = \sum_{j=1}^{J} \pi_j \, E(\underline{U}_j).$$

Hence $\text{cov}(\underline{U}) = E(\underline{U}\,\underline{U}^T) - E(\underline{U})[E(\underline{U})]^T$

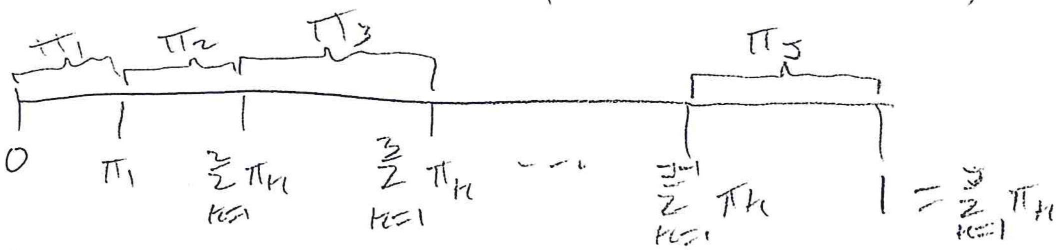
$$= \sum_{j=1}^{J} \pi_j \, E(\underline{U}_j \underline{U}_j^T) - E(\underline{U})[E(\underline{U})]^T =$$

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

If $E[\underline{U}_j] = \underline{0}$ for $j = 1, \ldots, J$, then $E(\underline{U}) = \underline{0}$ and $\text{cov}(\underline{U}) = \sum_{j=1}^{J} \pi_j \, \text{cov}(\underline{U}_j)$.

14) Random selection works: generate a Uniform(0,1) RV $W \perp\!\!\!\perp \underline{U}_j$ and set

$\underline{U} = \underline{U}_j$ if $W \in \left(\sum_{k=0}^{j-1} \pi_k, \sum_{k=0}^{j} \pi_k\right)$ with $\pi_0 = 0$.



15) Variable selection changes the dist of $\underline{U}_{jn}$ to $\underline{W}_{jn}$, say.

16) $\hat{\underline{\beta}}_{VS} = \hat{\underline{\beta}}_{I_k,0}$ with prob π_{kn}. Let $\hat{\underline{\beta}}_{MIX} = \hat{\underline{\beta}}_{I_k,0}$ with prob π_{kn} but $\hat{\underline{\beta}}_{MIX}$ uses random selection instead of variable selection.
Can't compute $\hat{\underline{\beta}}_{MIX}$ since the π_{kn} are unknown.

17) By OLS CLT, if $S \subseteq I_j$, then

$$\sqrt{n}\,(\hat{\underline{\beta}}_{I_j} - \underline{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\underline{0}, V_j).$$

Hence $\sqrt{n}\,(\hat{\underline{\beta}}_{I_j,0} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, V_{j0})$ where V_{j0} adds rows and columns of 0s corresponding to x_i not in I_j. Thus V_{j0} is singular unless I_j is the full model.

ex) $\underline{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (\beta_1, \beta_2, 0, 0)^T$ with

$\underline{\beta}_S = \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$, $\underline{\beta}_E = \begin{pmatrix}\beta_3\\ \beta_4\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}$. $\underline{\beta}_{I_1} = \underline{\beta}_S$, $a_1 = 2$

$\underline{\beta}_{I_2} = \begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\end{pmatrix}$, $a_2 = 3$, $\underline{\beta}_{I_3} = \begin{pmatrix}\beta_1\\ \beta_2\\ \beta_4\end{pmatrix}$, $a_3 = 3$ and $\underline{\beta}_{I_4} = \underline{\beta}$, $a_4 = 4$ are the I_j with $S = \{1, 2\} \subseteq I_j$