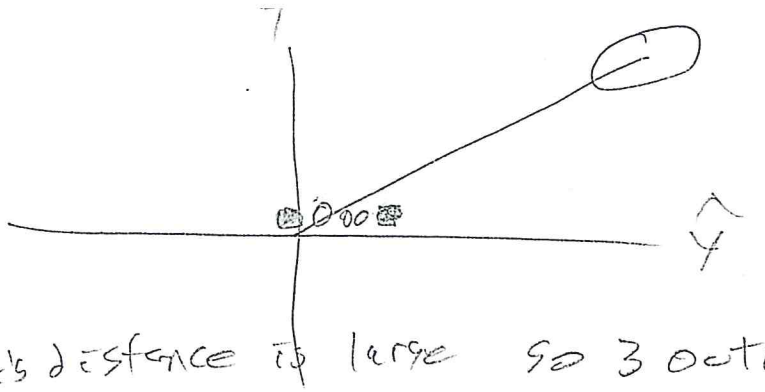
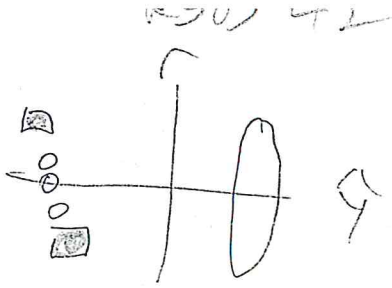
ex)



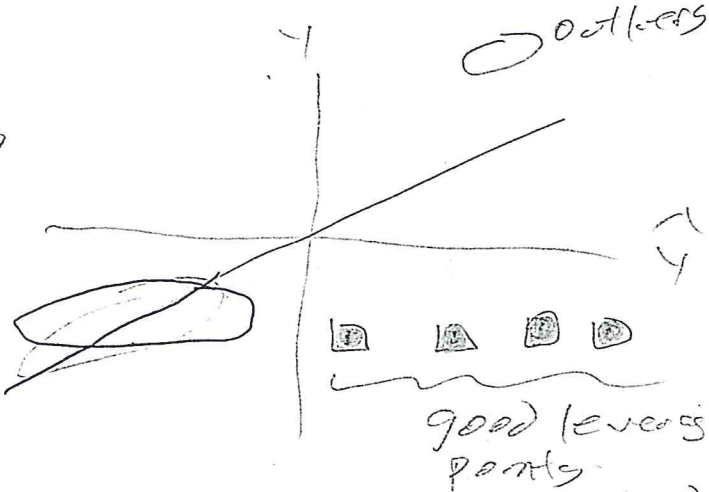


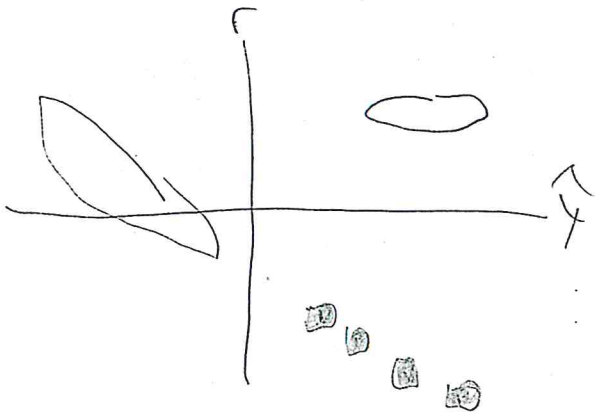
▣ = cook's distance is large so 3 outliers do not have large cook's distance. This is masking.. wrt cook's distance

None of the 5 outliers have large $|r_i|$; masking wrt residuals

All of the outliers have (small) $\hat{Y}_i$ (the 5 outliers are correctly detected by $\hat{Y}$).

ex)





The good leverage points had large cook's distances but the outliers did not. Masking for the outliers, swamping for the good leverage cases wrt cook's distances.

See HW 7 5.10.

ch6 practical resistant and Robust MLR Estimators

1) The MLD set MLR estimator $\hat{\beta}_D$ finds the coumbo set RMVN or RFCH set D applied to U_i = continuous predictors. Then $\hat{\beta}_D$ is the $\underset{MLR}{OLS}$ estimator.

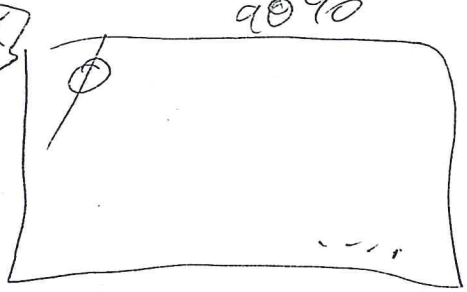
applied to the cases $(\underline{x}_i, y_i)$ corresponding to the $M \geq \frac{n}{2}$ cases in D. This estimator is √n consistent since the new responses Y_i were not used to select cases.

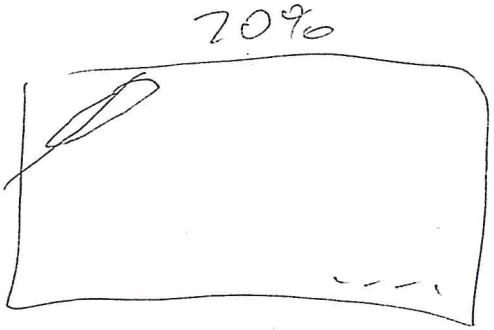
2) Compute (T, C) on the $\underline{x}_i$ (or $\underline{v}_i$).

Trim the M% cases with the largest D_i^2 and compute $\hat{\beta}_M$ from the remaining cases. Use $M = 0, 10, \ldots, 90$ to generate 10 response plots of $\hat{\beta}_M^T \underline{x}_i$ vs Y_i using all n cases.

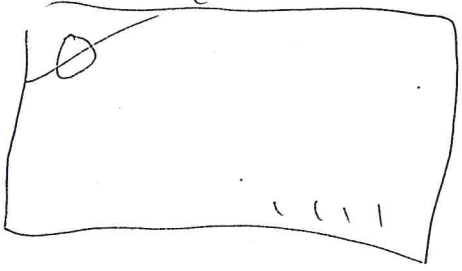
The trimmed views (TV or TVreg) estimator $\hat{\beta}_{T,n}$ corresponds to the response plot (trimmed view) where the width of the plotted points follow the identity line with the smallest variance function, ignoring any outliers.

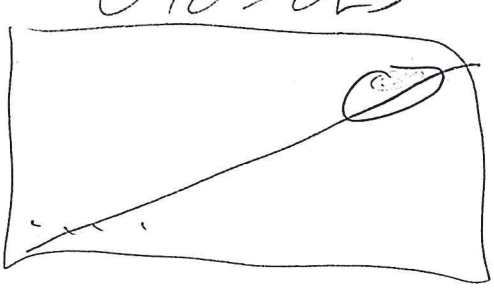
ex





$\hat{\beta}_{T,n}$ uses 70% trimming.





identity line goes through the outliers

bad

R58343

ex} data (0,1) (1,2) (2,3) (3,4) (4,11)
lie on $Y = 1 + X$

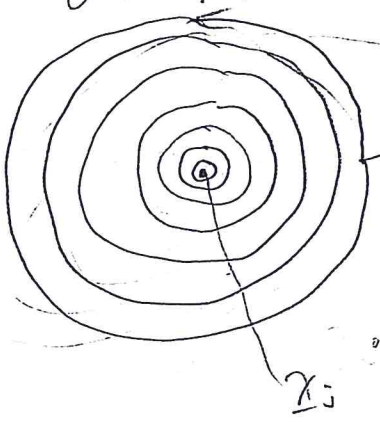
Let $J_1 = \{1,5\}$. The $p = 2^1$ cases are (0,1) and (1,11).

3} If $Y_i = \underline{x}_i^T \underline{\beta} + e_i$, $\hat{\underline{\beta}}$ is a resistant estimator if $\hat{\underline{\beta}}$ is known to be useful for detecting certain types of outliers.

ex} lmsreg and ltsreg in R are resistant estimators

4} Let $D_i(\underline{x}_j) = \sqrt{(\underline{x}_i - \underline{x}_j)^T (\underline{x}_i - \underline{x}_j)}$ be the Euclidean distance of $\underline{x}_i$ from $\underline{x}_j$ for $i = 1, \ldots, n$. Let $D_{(1)}(\underline{x}_j) \leq \cdots \leq D_{(n)}(\underline{x}_j)$ be the order statistics.

5} For a fixed $\underline{x}_j$ consider the cases $(Y_i, \underline{x}_i)$ corresponding to the α percentage of $\underline{x}_i$ closest to $\underline{x}_j$ with $\alpha \in \{1, 2.5, 5, 10, 20, 33.3\}$,



$\left\{ \underline{z} \mid (\underline{z} - \underline{x}_j)^T (\underline{z} - \underline{x}_j) \leq D^2_{\left(\left\lceil \frac{\alpha n}{100} \right\rceil\right)} \right\}$

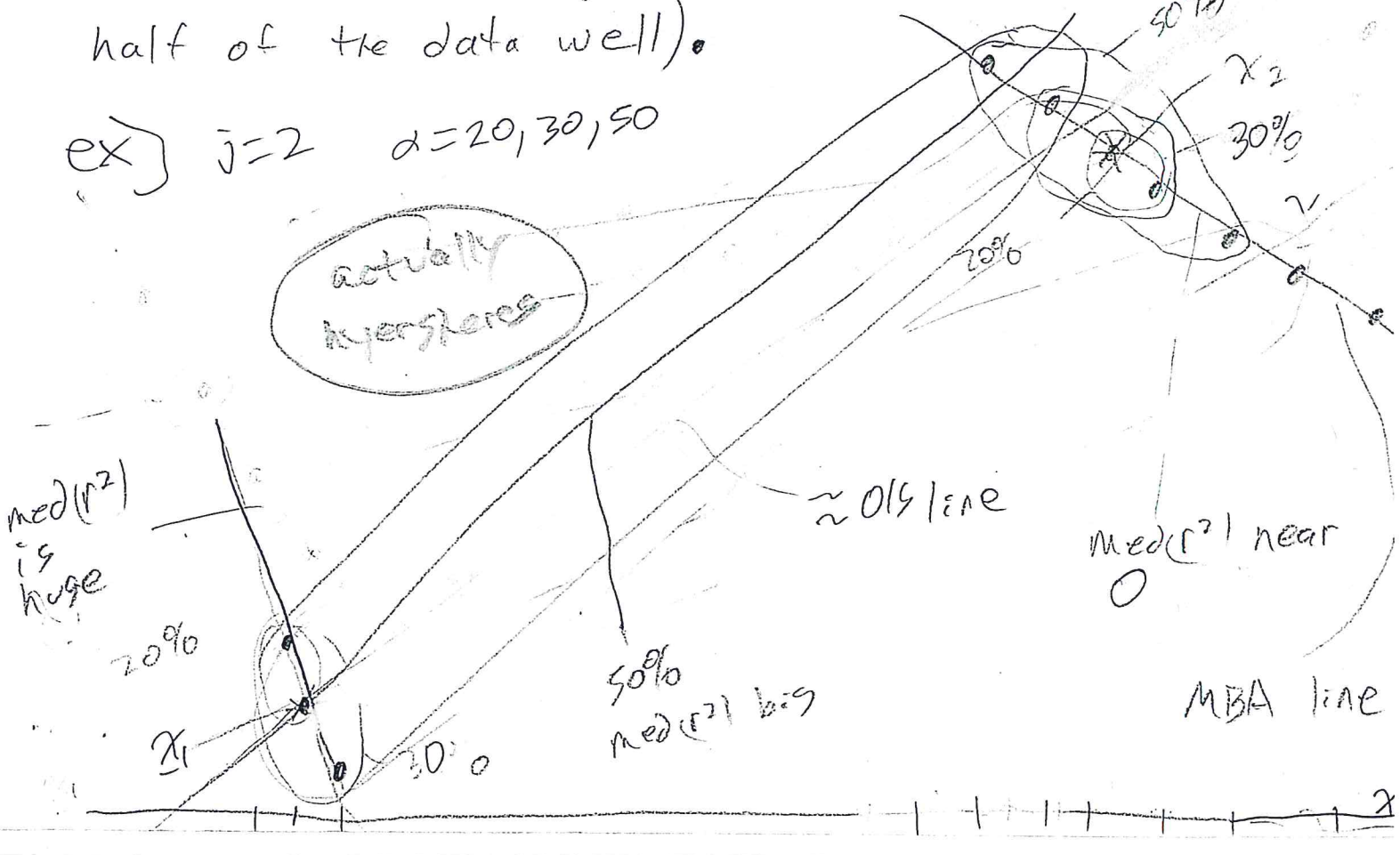
if $\alpha = 50$ about 50% of the $\underline{x}_i$ are in the sphere

$\hat{\beta}_{MBA}$ is a resistant estimator. Randomly select p x_j from the n. Let $\hat{\beta}_{\alpha j}$ denote the OLS fit to the $\alpha\%$ cases with x_i closest to x_j for $\alpha \in \{1, 2.5, 5, 10, 20, 33, 50\} = A$. Also compute $\hat{\beta}$, the OLS fit to all of the data. This yields $50p$ OLS fits $\hat{\beta}_{\alpha j}$ for $\alpha \in A$ and $j = 1, \ldots, p$ and $\hat{\beta}$ $(= \hat{\beta}_{100,j} = \hat{\beta}_{100})$. Compute the median squared residual

$$\text{med}\left((y_i - x_i^T \hat{\beta}_{\alpha j})^2\right)$$

for each of the $50p$ fits. $\hat{\beta}_{MBA}$ corresponds to the $\hat{\beta}_{\alpha j}$ with the smallest median squared residual (so fits half of the data well).

ex) $j = 2$ $\alpha = 20, 30, 50$



7) The rpmreg2 estimator $\hat{\beta}_R$ is the OLS estimator applied to the RMVN set computed from $\underline{U}_i = (x_{i2}, \ldots, x_{ip}, Y_i)^T$ where $x_{i1} \equiv 1$. $\hat{\beta}_R$ is $\sqrt{n}$ consistent if the $\underline{U}_i$ come from a large class of EC dist'ls which is a much stronger assumption than i.i.d errors e_i.

§6.2

8) An estimator $\hat{\beta}$ is high breakdown (HB) if the median absolute or squared residual stays bounded even if nearly half (but $< \frac{n}{2}$) of the cases are contaminated.

9) If $\hat{\beta}_0$ is high breakdown, so is the attractor after k FLTS concentration steps.

10) If a constant is in the model $\hat{\beta}_M = (\overbrace{MED(n)}^{\text{of the } Y_i}, 0, \ldots, 0)^T$ has median absolute residual $= MAD(n)$. Hence $\hat{\beta}_M$ is HB. Also $\|\hat{\beta}_M\| = |MED(n)|$ and $MED(n)$ is HB.

11) Let $p_M = p - 1$. Let the OLS fit $\underline{\beta}_0$ to the cases whose Y values are closest to MED(Y), let $\hat{\underline{\beta}}_B$ be the attractor after k concentration steps.
Let $\underline{b}_{kB} = 0.9999\, \hat{\underline{\beta}}_B$. Then $\underline{b}_{kB}$ and $\hat{\underline{\beta}}_B$ are HB MLR estimators.

12) The lbreg estimator $\hat{\underline{\beta}}_H$ uses 3 attractors.

i) $\hat{\underline{\beta}}_C = \hat{\underline{\beta}}_{OLS}$ a consistent estimator

ii) $\hat{\underline{\beta}}_A = \hat{\underline{\beta}}_{MBA}$ (default) or $\hat{\underline{\beta}}_{rmreg2}$ or practical other resistant estimator.

iii) $\hat{\underline{\beta}}_B$, a practical HB attractor.

Pick $a = 1.4 > 1$ and set $\hat{\underline{\beta}}_H = \hat{\underline{\beta}}_C$. If $a\,Q_L(\hat{\underline{\beta}}_A) < Q_L(\hat{\underline{\beta}}_C)$, set $\hat{\underline{\beta}}_H = \hat{\underline{\beta}}_A$. If $a\,Q_L(\hat{\underline{\beta}}_B) < \min[Q_L(\hat{\underline{\beta}}_C), a\,Q_L(\hat{\underline{\beta}}_A)]$, set $\hat{\underline{\beta}}_H = \hat{\underline{\beta}}_B$. Q_L is the LMS, LTA (default) or LTS criterion.

13) Let $\hat{\underline{\beta}}_L$ be the LMS, LTS or LTA estimator. If $\hat{\underline{\beta}}_C$ and $\hat{\underline{\beta}}_L$ are both consistent estimators of $\underline{\beta}$, then $\hat{\underline{\beta}}_H$ is a HB MLR estimator asymptotically equivalent to $\hat{\underline{\beta}}_C$.

14) Typically $\hat{\underline{\beta}}_C$ and $\hat{\underline{\beta}}_L$ are both consistent estimators for $\underline{\beta}$ if the iid e_i are from a large class

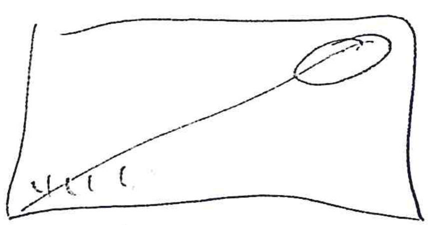
of zero mean finite variance symmetric distributions. (K>0 ? 4)

If the distribution is skewed the constant β_1 estimated by $\hat{\beta}_L$ and $\hat{\beta}_C$ are different. Then $\hat{\beta}_H$ tends to select $\hat{\beta}_B$ too often.

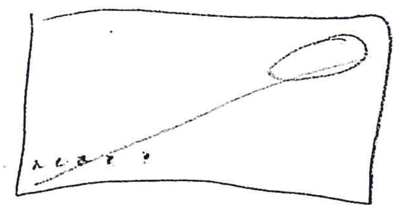
15) $\hat{\beta}_B$ tends to be poor for e_i symmetric. If the e_i dist is skewed, $\hat{\beta}_B$ seems to estimate the slopes ($\hat{\beta}_i$ $i > 1$) fairly well.

16) Just as the MCD criterion sometimes selected (T_{DGK}, C_{DGK}) instead of (T_{MB}, C_{MB}) when outliers make (T_{DGK}, C_{DGK}) bad, the LTA criterion sometimes select $\hat{\beta}_C$ when the response plots look as follows:

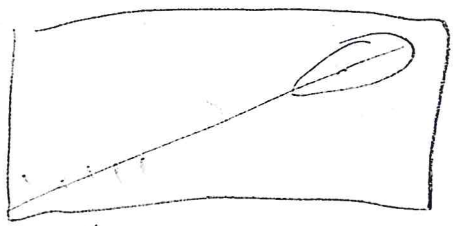
4



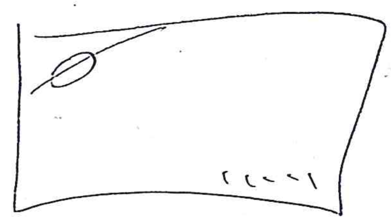
HB FIT



OLS FIT



$\hat{\beta}_A$ FIT



BB FIT $\hat{\beta}_B$

end exam 2 material

ch 7 Variable Selection and Lasso (454)

1) variable selection is a search for a subset of predictors that can be deleted without important loss of information if $n \geq 10p$ or so that model I is good for prediction if $n < 5p$.
I = remaining predictors in the model

2) Let I_{min} be the model selected by the variable selection method. (for forward selection and backward elimination) use the model that minimizes $C_p(I)$ if $n \geq 10p$ $EBIC(I)$ if $n < 10p$.

3)* A model for variable selection is

$$\underline{x}^T\underline{\beta} = \underline{x}_S^T\underline{\beta}_S + \underline{x}_E^T\underline{\beta}_E = \underline{x}_S^T\underline{\beta}_S$$

where $\underline{x} = (\underline{x}_S^T\ \underline{x}_E^T)^T$, $\underline{x}_S$ is $a_S \times 1$, and $\underline{x}_E$ is $(p-a_S)\times 1$. Given that $\underline{x}_S$ is in the model, $\underline{\beta}_E = \underline{0}$ and E denotes the subset of terms that can be eliminated given S is in the model. (S is the model you would like to select.)

4) Let $\underline{x}_I$ denote an $a \times 1$ vector of terms in the candidate model. If $S \subseteq I$ and 3)

holds, then $\underline{x}^T\underline{\beta} = \underline{x}_S^T\underline{\beta}_S = \underline{x}_I^T\underline{\beta}_I + \underline{x}_0^T\underline{\beta}_0 = \underline{x}_I^T\underline{\beta}_I$.

($\underline{x}_I$: in model; $\underline{x}_0$: out of model)

ex) * $p=4$ a constant β_1 is always in the model

$\underline{\beta} = (\beta_1, \beta_2, 0, 0)^T$ There are $J = 2^{p-1} = 8$ possible subsets of $\beta_1, \ldots, \beta_3$ that always contain 1 are $I_1 = \{1\}$, $S = I_2 = \{1,2\}$, $I_3 = \{1,3\}$, $I_4 = \{1,4\}$, $I_5 = \{1,2,3\}$, $I_6 = \{1,2,4\}$, $I_7 = \{1,3,4\}$ and $I_8 = \{1,2,3,4\}$. There are $2^{p-q} = 4$ subsets I_2, I_5, I_6 and I_8 such that $S \subseteq I_j$. Let $\hat{\underline{\beta}}_{I_7} = (\hat{\beta}_1, \hat{\beta}_3, \hat{\beta}_4)^T$ and $\underline{x}_{I_7} = (x_1, x_3, x_4)^T$.

ex) * If $\underline{\beta}_I$ is $q \times 1$, use zero padding to form a $p \times 1$ vector $\hat{\underline{\beta}}_{I,0}$.

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

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

I_j	model	x_2	x_3	x_4	x_5	$\hat{\beta}_{I_j,0}$
I_2	1		*			$(\hat{\beta}_1, 0, \hat{\beta}_3, 0, 0)^T$
I_3	2		*	*		$(\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4, 0)^T$
I_4	3	*	*	*		$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, 0)^T$
I_5	4	*	*	*	*	$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)^T$

full model

Assume $X_1 \equiv 1$ is always in the model and $p = 5$.

Given $C_p(I_j)$ if $I_{min} = I_3$

Then $\hat{\beta}_{US} = \hat{\beta}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4, 0)^T$.

6) $Y = \underline{x}^T \underline{\beta} + e$ = full model

$Y = \underline{x}_I^T \underline{\beta}_I + e$ = submodel.

The full model is a submodel.

7) Underfitting occurs if $S \not\subseteq I$ so $\underline{x}_I$ is missing important predictors. Underfitting occurs if $\underline{x}_I$ is $a \times 1$ with $a < a_S$ where $\underline{\beta}_S$ is $a_S \times 1$. Overfitting occurs if $n < 5a$ or if $S \subseteq I$ but $S \neq I$.

$n < 5a$: not enough data to estimate a parameters well.

Overfitting is serious if $n < 5a$, but "not much of a problem" if $n > 10p$ or $n > 20p$.