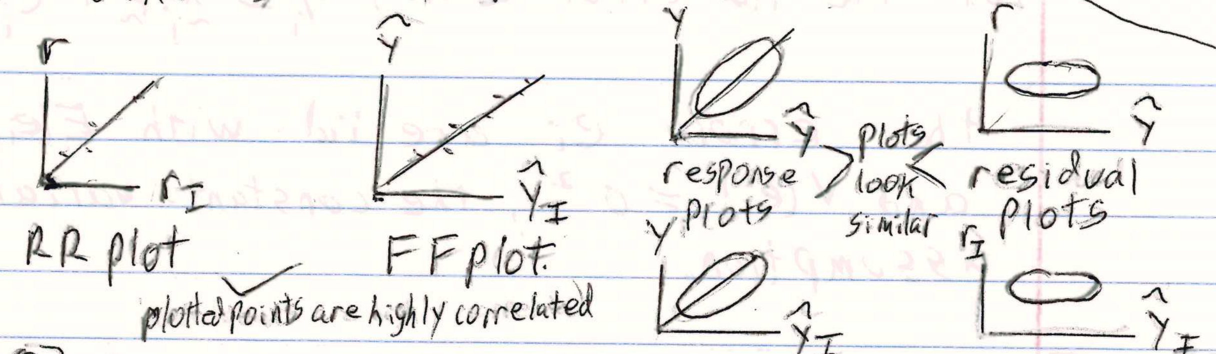
67] p105 * Full model and candidate for final submodel I



plotted points are highly correlated

68] p104 For models with k predictors, the model with the smallest $C_p(I)$ maximizes $corr(r, r_I)$, $C_p(I) \le 2k \Rightarrow corr(r, r_I) \ge \sqrt{1 - \frac{p}{n}}$, and as $corr(r, r_I) \to 1$, $corr(\hat{Y}, \hat{Y}_I) \to 1$.

69] Prop 3.2 proves that for models with a constant, the plotted points scatter about the identity line for RR, FF and response plots.
Go to 83)–86).

70] p144 Let R_k^2 be the R^2 from regressing x_k on the other predictors. If R_k^2 is high, multicollinearity is present and $SE(\hat{\beta}_k) \propto \frac{1}{1-R_k^2}$ is large. Highly correlated linear plots in the scatterplot matrix also suggest multicollinearity. Multicollinearity makes CI for β_k very long, $(X^T X)$ nearly singular, but is "not a problem for predicting Y."

§3.5 71] p129 Diagnostics are used to

Check that model assumptions are reasonable. (36.5

For the iid error model, $Y_i = \underset{\sim}{x}_i^T \underset{\sim}{\beta} + e_i$,

the errors e_i are iid with $E(e_i) = 0$ and $V(e_i) = \sigma^2$, the constant variance assumption.

72] p 132 The best diagnostics are the response and residual plots, but numerical diagnostics are popular.

73] In matrix form $\underline{Y} = \underline{X} \underset{\sim}{\beta} + \underset{\sim}{e}$

$$\underset{n \times p}{\underline{X}} = [x^1, x^2, \ldots, x^p] = [\underset{\sim}{1}, U]$$

where $\underset{n \times (p-1)}{U} = [x^2, \ldots, x^p]$ is the matrix of

nontrivial predictors. Let $\underset{\sim}{x}_i = (1, x_{i2}, \ldots, x_{ip})^T = (1, \underline{u}_i^T)^T$ be the ith case, the

ith row of $\underline{X}$.

74] p13D The sample mean and covariance matrix of the nontrivial predictors $(\underline{u}_1, \ldots, \underline{u}_n)$ are $\underset{\sim}{\overline{u}} = \frac{1}{n} \sum_{i=1}^{n} \underset{\sim}{u}_i$

and $C = \widehat{Cov}(U) = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{u}_i - \underset{\sim}{\overline{u}})(\underline{u}_i - \overline{u})^T$.

$[\widehat{Cov}(U)]^{-1}$ is closely related to $(\underline{X}^T\underline{X})^{-1}$.

75] $\hat{\underline{Y}} = H \underset{\sim}{Y}$, $H = \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T$

76] Leave one out or case diagnostics are computed by omitting the ith case.

Let $\hat{\underset{\sim}{Y}}_{(i)} = \underline{X}\, \hat{\underset{\sim}{\beta}}_{(i)}$ be the fitted values when the ith case is omitted.

$\leftarrow h_i \le 1 \quad V(r_i) \ge 0$

$V(r_i) = \sigma^2 (1-h_i)$, $h_i = h_{ii} = H_{ii}$, $0 \le h_i \le 1$, is the ith <u>leverage</u>. The <u>studentized residual</u> $\hat{e}_i = \dfrac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$ where $\hat{\sigma}^2 = MSE$.

77] * p130-1 $MD_i^2 = (\underset{\sim}{u}_i - \bar{\underset{\sim}{u}})^T C^{-1} (\underset{\sim}{u}_i - \bar{u})$ is the ith squared <u>Mahalanobis distance</u>.

The ith <u>Cook's distance</u>

$$CD_i = \frac{(\hat{\underset{\sim}{Y}}_{(i)} - \hat{\underset{\sim}{Y}})^T (\hat{\underset{\sim}{Y}}_{(i)} - \hat{\underset{\sim}{Y}})}{p\, \hat{\sigma}^2}$$

$$= \frac{r_i^2}{p\hat{\sigma}^2 (1-h_i)} \frac{h_i}{1-h_i} = \frac{\hat{e}_i^{\,2}}{p} \underbrace{\frac{h_i}{1-h_i}}_{0 < h_i \le 1}.$$

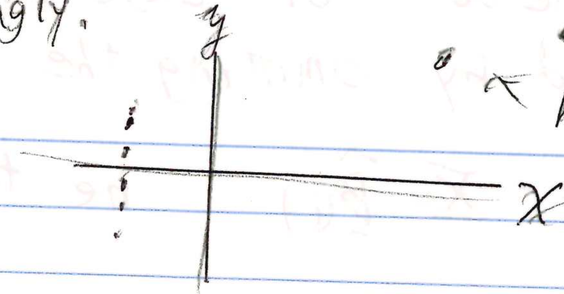
CD_i measures "influence" of case i on $\hat{\underset{\sim}{\beta}}$.

influence ≈ (leverage) (distance of Y_i from $\hat{Y}_i$)

78] Leverage h_i is large if $\underset{\sim}{u}_i$ is far from $\bar{\underset{\sim}{u}}$ where "far" depends on $[\widehat{Cov}(\underset{\sim}{u})]^{-1}$ or $(\underline{X}^T\underline{X})^{-1}$.

$\sum_{i=1}^n h_{ii} = p$, $0 < h_i \le 1$, $h_i = \underset{\sim}{x}_i^T (\underline{X}^T\underline{X})^{-1} \underset{\sim}{x}_i$ and $h_i = \frac{1}{n-1} MD_i^2 + \frac{1}{n}$.

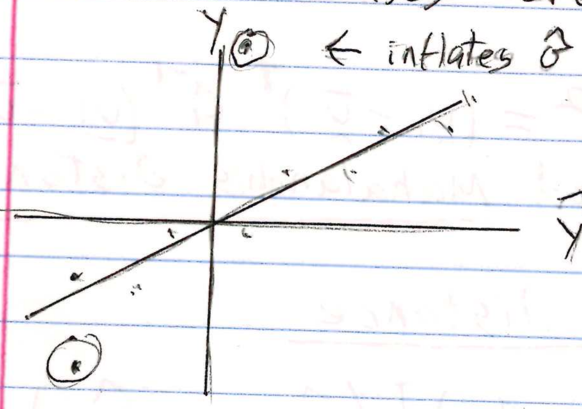
Cases with high h_i and CD_i tend to effect $\hat{\beta}$ strongly. 37.5



← may have $r_i \approx 0$ and low CD_i

← high leverage since $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ changes as this point is moved vertically.

79] In the response plot, cases with large CD_i tend to have large $|r_i|$ if they are close to the middle of the plot. The $|r_i|$ need not be so large if the cases are far from the middle.



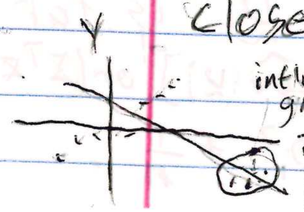
← inflates $\hat{\sigma}$ but does not effect $\hat{\beta}$ much, could have low CD_i since h_i is low: near 0

80] $h_i > \frac{2p}{n}$, $MD_i^2 > \chi^2_{p-1, .95}$,

$CD_i > \min\left(.5, \frac{2p}{n}\right)$ are large

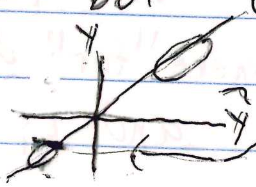
and should be checked.

81] These numerical diagnostics do not work well if 2 or more cases are close together but unusual.



influential group on $\hat{\beta}$

most cases in group have small h_i, CD_i



many cases will not have large CD_i, h_i or MD_i