$\hat{Y}_f$ depends on $Y_1, \ldots, Y_n$ so $\hat{Y}_f \perp\!\!\!\perp Y_f$

$$= V(\underline{x}_f' \hat{\underline{\beta}}) + \sigma^2 = V(\underline{x}_f' (\underline{X}'\underline{X})^{-1} \underline{X}' \underline{Y}) + \sigma^2$$

$$= \left[\underline{x}_f' (\underline{X}'\underline{X})^{-1} \underline{X}' \sigma^2 \underline{I} \, \underline{X} (\underline{X}'\underline{X})^{-1} \underline{x}_f \right] + \sigma^2$$

$$= \sigma^2 \Big(\underbrace{\underline{x}_f' (\underline{X}'\underline{X})^{-1} \underline{x}_f}_{h_f} + 1 \Big) = \sigma^2 (1 + h_f)$$

p132

16] If the ε_i are iid $N(0, \sigma^2)$, then a $100(1-\delta)\%$ prediction interval (PI) for the random variable Y_f is

$$\hat{Y}_f \pm t_{n-p, 1-\frac{\delta}{2}} \sqrt{MSE} \sqrt{1 + h_f}$$ (closed interval)

want $h_f \le \max h_i$, $h_i = \underline{x}_i' (\underline{X}'\underline{X})^{-1} \underline{x}_i$

= ith diagonal entry of $H = P = \underline{X} (\underline{X}'\underline{X})^{-1} \underline{X}'$

17] As $n \to \infty$ the PI in 16] estimates

$\left[E(Y_f | \underline{x}_f) - z_{1-\frac{\delta}{2}} \sigma, \; E(Y_f | \underline{x}_f) + z_{1-\frac{\delta}{2}} \sigma \right]$, the highest $1-\delta$ density region if

$Y_f | \underline{x}_f \sim N(E(Y_f|\underline{x}_f), \sigma^2)$.

18] KNOW A large sample $100(1-\delta)\%$ PI

$[L_n, U_n]$ satisfies $P[Y_f \in [L_n, U_n]] \to 1-\delta$

as $n \to \infty$. actually $Y_f|\underline{x}_f$ but suppress $\underline{x}_f$

19] Suppose $Y_f | \underline{x}_f$ (Y_f) has pdf $f_f(y)$ and cdf $F_f(y)$.

Want $[L_n, U_n] \xrightarrow{P} [L, U]$ where $F_f(U) - F_f(L) = 1-\delta$

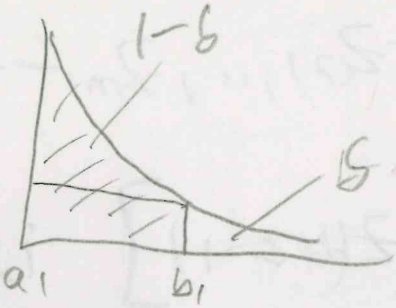
and $U-L$ is short,

20] The highest density region is found by moving a horizontal line down from the top of the pdf. The line will intersect the pdf or boundaries of the support of the pdf at $[a_1, b_1], \ldots, [a_H, b_H]$ for some $H \geq 1$. Stop moving the line when the areas under the pdf corresponding to the intervals is $1-\delta$. Often the pdf is unimodal and decreases rapidly as y moves away from the mode. Then $K=1$ and the highest density region is an interval.

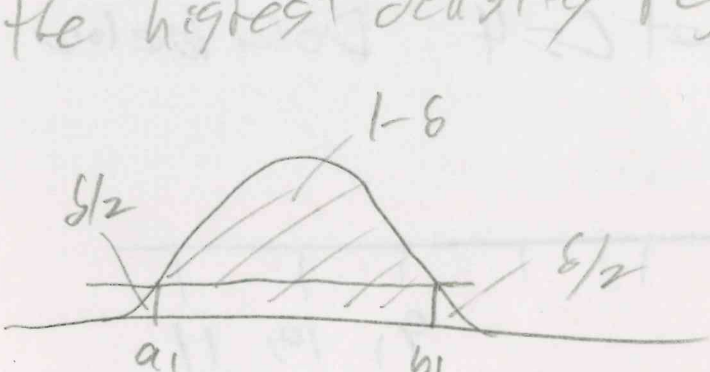
ex) a) If Y_t has an exponential distribution, then the highest density region is

$[0, \xi_{1-\delta}]$ where $P(Y_t \le \xi_\alpha) = \alpha$

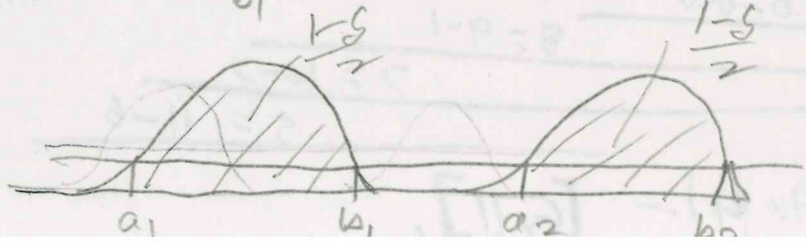
xi (zi)



b) For a symmetric unimodal distribution, the highest density region is $[\xi_{\delta/2}, \xi_{1-\delta/2}]$.



c)



bimodal pdf

2B) Suppose you have data

$z_1, \ldots, z_n$. The order

statistics $z_{(1)} \le z_{(2)} \le \cdots \le z_{(n)}$

$z_{(1)} = \min(z_1, \ldots, z_n)$

$z_{(n)} = \max(z_1, \ldots, z_n)$,

Consider intervals that contain c

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

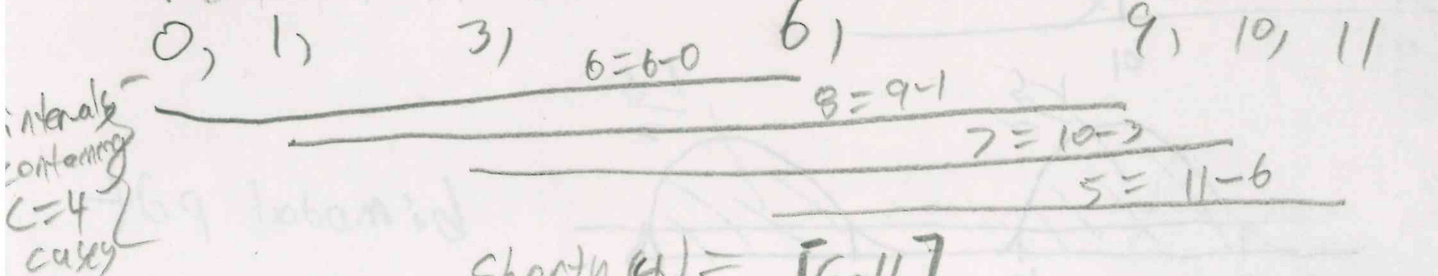
Compute $z_{(c)} - z_{(1)}, z_{(c+1)} - z_{(2)}, \ldots, z_{(n)} - z_{(n-c+1)}$,

Then $\text{Shorth}(c) = [z_{(d)}, z_{(d+c-1)}]$ is

the closed interval with the shortest length.

ex) know for quiz Let $c=4$ Data below

has $n=7$.



Shorth(4) = [6, 11].

22) If $Y_1, \ldots, Y_n$ are iid

$\underline{Y} = \underset{\sim}{1} \beta_0 + \underset{\sim}{e}$ and $\frac{c}{n} \to 1-\delta$

eg $c = k_n = \lceil n(1-\delta) \rceil$, then the shorth (c) interval estimates the highest density $100(1-\delta)\%$ region if that region is an interval. Then the "shorth(c)" interval can be used as a large sample $100(1-\delta)\%$ PI for Y_f. If $c = k_n$ then for large $n\delta$ and iid data, the shorth PI has max. undercoverage $\approx 1.12\sqrt{\delta/n}$. So using

$c = \lceil n [1-\delta + 1.12\sqrt{\frac{\delta}{n}}] \rceil$ works better than $c = k_n$. (Frey 2013)

23) Let $a_n = (1 + \frac{15}{n})\sqrt{\frac{n}{n-p}}\sqrt{1+h_f}$

Let $c = k_n$ and find the shorth estimator applied to the residuals $e_1, \ldots, e_n$.

so shorth(c) = $[e_{(d)}, e_{(d+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$.

Let $\underline{Y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}$, $E\underline{\varepsilon} = 0$, $Cov(\underline{\varepsilon}) = \sigma^2 I$.

Then a large sample $100(1-\delta)\%$ PI for

Y_f is $\left[\hat{Y}_f + a_n \tilde{\xi}_{\delta_1},\ \hat{Y}_f + a_n \tilde{\xi}_{1-\delta_2}\right]$.

This PI is asymptotically optimal (short) if the $\underline{x}_i$ are bounded in probability and the iid ε_i come from a large class of zero mean distributions.

---

skip § 5.3.2, 5.4

---

ch 9 1) $\underline{Y} = \underline{X}\,\underline{\beta} + \underline{\varepsilon}$ $\quad$ $\underline{X}$ full rank

$E(\underline{\varepsilon}) = \underline{0}$ $\quad$ $Cov(\underline{\varepsilon}) = \sigma^2 I$.

We have i) assumed the model is correct, at least approximately. In practice this assumption is often violated.

2} i) could have $E\underline{Y} \neq \underline{X}\underline{\beta}$ if

$\underline{X}$ is missing important predictor variables.

ii) could have $\operatorname{cov}(\underline{\varepsilon}) \neq \sigma^2 \underline{I}$

iii) The ε_i could be correlated instead of iid (uncorrelated).

§ 9.2 3} Suppose we fit model

$\underline{Y} = \underline{X}\underline{\beta} + \underset{\sim}{\varepsilon}$, but the true model

is $\underline{Y} = \underline{X}\underline{\beta} + \underline{W}\underset{\sim}{\gamma} + \underline{\varepsilon}$

$\uparrow$

columns of $\underline{W}$ should be in the model

Where the columns of the $n \times q$ full rank matrix $\underline{W}$ are linearly independent of the columns of the full rank $n \times p$ matrix $\underline{X}$.

Then $E\hat{\underline{\beta}} = E[(\underline{X}'\underline{X})^{-1}\underline{X}'\underline{Y}] =$

$(\underline{X}'\underline{X})^{-1}\underline{X}'[\underline{X}\underline{\beta} + \underline{W}\underline{\gamma}] = \underset{\sim}{\beta} + (\underline{X}'\underline{X})^{-1}\underline{X}'\underline{W}\gamma$

$= \underset{\sim}{\beta} + L\underset{\sim}{\gamma}$. So $\hat{\underline{\beta}}$ is a biased estimator of $\underset{\sim}{\beta}$ with bias $L\underset{\sim}{\gamma}$. This bias could be quite large, but $L\underset{\sim}{\gamma} = \underset{\sim}{0}$ if $\underset{\sim}{X}'\overline{W} = 0$ ie if the columns of $\overline{W}$ are orthogonal to the columns of $\underset{\sim}{X}$.

4) Leaving out important predictors can destroy the linearity of the model

ex) Fit $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ when the true model is $Y_i = \beta_0 + \beta_1 \underbrace{x_{i1}}_{x_{i1} = x_i} + \beta_2 \underbrace{x_{i1}^2}_{x_{i2} = x_{i1}^2} + \varepsilon_i$

cook and weisberg P264-5

ex 5) Suppose $\underset{\sim}{x} = \begin{pmatrix} x_1 \\ x_{p-1} \end{pmatrix}$, $\underset{\sim}{\beta} = \begin{pmatrix} \beta_1 \\ \beta_{p-1} \end{pmatrix}$, and

$E(Y \mid \underset{\sim}{x}) = \underset{\sim}{\beta_1}^T \underset{\sim}{x_1} + \beta_{p-1} x_{p-1} = \underset{\sim}{\beta}^T \underset{\sim}{x}$.

consider regressing Y on $\underset{\sim}{x_1}$, so without x_{p-1}.

Then $E(Y|\underline{x}_1) = E\left[E(Y|\underline{x}_1, x_{p-1})|\underline{x}_1\right]$

nontrivial fact: $EW = E(E[W|x_{p-1}])$

but $E(W|x_{p-1}) = E(Y|\underline{x}_1, x_{p-1})$ take $W = Y|\underline{x}_1$

$= E\left(\underline{\beta}_1^T \underline{x}_1 + \beta_{p-1} x_{p-1} \,|\, \underline{x}_1\right)$

$= \underline{\beta}_1^T \underline{x}_1 + \beta_{p-1} \underbrace{E(x_{p-1}|\underline{x}_1)}$

x_{p-1} gets replaced by $E(x_{p-1}|\underline{x}_1)$ when x_{p-1} is omitted from the LS model.

ex) $E(Y|\underline{x}) = 1 + 2x_1 + 3x_2$, $V(Y|\underline{x}) = \sigma^2$.

Then $E(Y|x_1) = 1 + 2x_1 + 3E(x_2|x_1)$

a) If $x_1 \perp\!\!\!\perp x_2$ then $E(x_2|x_1) = E(x_2)$

and $E(Y|x_1) = (1 + 3E(x_2)) + 2x_1$

is linear. The coefficient for x_1 does not change but the intercept does.

b) Suppose $E(x_2|x_1) = \alpha_0 + \alpha_1 x_1$

Then $E(Y|x_1) = 1 + 2x_1 + 3\alpha_0 + 3\alpha_1 x_1$

$= (1+3\alpha_0) + (2+3\alpha_1)x_1$ which is again linear, but both the intercept and slope have changed.

c) If $E(x_2|x_1) = \alpha_0 + \alpha_1 \exp(\alpha_2 x_1)$, then

$$E(Y|x_1) = 1+2x_1 + 3\alpha_0 + 3\alpha_1 \exp(\alpha_2 x_1)$$

$$= (1+3\alpha_0) + 2x_1 + 3\alpha_1 \exp(\alpha_2 x_1)$$

which is a nonlinear mean function.

6) under the conditions of 5],

$$V(Y|\underline{x_1}) = E\Big[V(\overbrace{Y|\underline{x_1}, x_{p-1}}^{\text{correct linear model}})|\underline{x_1}\Big] + V\Big[E(Y|\underline{x_1}, x_{p-1})|\underline{x_1}\Big]$$

$$= E(\sigma^2|\underline{x_1}) + V\Big[(\underbrace{\underline{\beta_1}^T \underline{x_1}}_{\text{constant given } \underline{x_1}} + \beta_{p-1}x_{p-1})|\underline{x_1}\Big]$$

$$= \sigma^2 + \beta_{p-1}^2 V(x_{p-1}|\underline{x_1}).$$

Hence deleting a term from the model may result in a nonconstant variance function.

For a linear model when x_{p-1} is omitted,

7) want $E(x_{p-1} | \underline{x}_1) = \underline{\gamma}^T \underline{x}_1$

and $V(x_{p-1} | \underline{x}_1) = \tau^2$

so $E(Y | \underline{x}_1) = \underset{\sim}{\beta}_1^T \underline{x}_1 + \beta_{p-1} \underline{\gamma}^T \underline{x}_1 = \underline{\eta}^T \underline{x}_1$

$\underline{\eta} = \underset{\sim}{\beta}_1 + \beta_{p-1} \underline{\gamma}$

and $V(Y | \underline{x}_1) = \sigma^2 + \beta_{p-1}^2 V(x_{p-1} | \underline{x}_1) =$

$\sigma^2 + \beta_{p-1}^2 \tau^2 = \theta^2$ say.

8) If $\underline{x} = (1, \underbrace{x_2, \ldots, x_{p-1}}_{\underset{\sim}{w}'})' = (1 \ \ \underline{w}')'$

and $\begin{pmatrix} Y \\ x_2 \\ \vdots \\ x_{p-1} \end{pmatrix} \sim N_p\left(\underline{\mu}, \begin{pmatrix} \sigma_Y^2 & \Sigma_{Y\underline{w}} \\ \Sigma_{\underline{w}Y} & \Sigma_{\underline{w}} \end{pmatrix}\right),$

then $Y | x_{i1}, \ldots, x_{ik}$ follows a linear model with constant variance

$Y_i = \beta_{0k} + \beta_{1k} x_{i1} + \cdots + \beta_{kk} x_{ik} + \varepsilon_{ik}$

$V(\varepsilon_{ik}) = \sigma_k^2$. Models with lower σ_k^2 are better.

9) Having too many predictors (if $n \geq 10p$) is much less serious than having too few, since the $\hat{\beta}_i$ for unneeded x_i tend to have $\hat{\beta}_i \xrightarrow{P} 0$.

10) p230 overfitting: Suppose $E(\underline{Y}) = \underset{n \times k}{\underline{X}_1} \underline{\beta}_1$ where

$$\underline{X} = (\underline{X}_1 \ \underline{X}_2) \text{ and } \underline{\beta} = \begin{pmatrix} \underline{\beta}_1 \\ \underline{0} \end{pmatrix} \text{ since } \underline{\beta}_2 = \underline{0}.$$

Then $\underline{X}_1 \underline{\beta}_1 = \underline{X}\,\underline{\beta} = (\underline{X}_1 \ \underline{X}_2) \begin{pmatrix} \underline{\beta}_1 \\ \underline{0} \end{pmatrix}$,

So $E(\hat{\underline{\beta}}) = (\underline{X}'\underline{X})^{-1} \underline{X}' \underline{X}_1 \underline{\beta}_1 =$

$$(\underline{X}'\underline{X})^{-1} \underline{X}' \underline{X} \begin{pmatrix} \underline{\beta}_1 \\ \underline{0} \end{pmatrix} = \begin{pmatrix} \underline{\beta}_1 \\ \underline{0} \end{pmatrix} = \underline{\beta}.$$

Also $E\,\hat{\underline{Y}} = E(\underline{X}\hat{\underline{\beta}}) = \underline{X} \begin{pmatrix} \underline{\beta}_1 \\ \underline{0} \end{pmatrix} = \underline{X}\,\underline{\beta} = \underline{X}_1 \underline{\beta}_1$

and $E(MSE) = \sigma^2$.

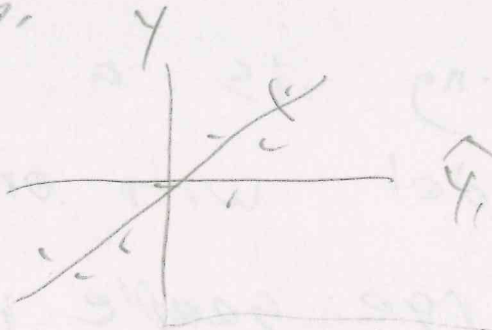
However R^2 is too high and the 1st k diagonal elements of $(\underline{X}'\underline{X})^{-1}$

are larger than the diagonal elements of $(X_1'X_1)^{-1}$. So CIs for β_i using X are longer than the CIs for β_i using X_1 for $i = 1, \ldots, k$.

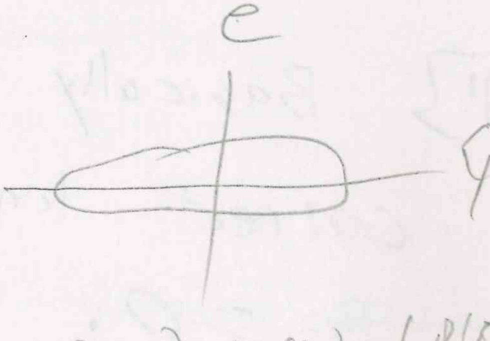
11) Basically overfitting is a correct linear model with one or more $\beta_i = 0$. So large sample inference is correct but not as precise as using the model that omits the predictors with $\beta_i = 0$.
want $n \geq 10k$ if $Y = \underset{n \times k}{X_1}\beta_1 + e$, $n \geq 10p$ if $\underline{Y} = \underset{n \times p}{X}\beta + e$

§ 9.3 12) Suppose $\underline{Y} = X\underline{\beta} + \underline{e}$, $E(\underline{e}) = \underline{0}$ but $Cov(\underline{e}) = \sigma^2 V$ instead of $\sigma^2 I_n$. Then $\hat{\underline{\beta}} \xrightarrow{P} \underline{\beta}$ but $Cov(\hat{\underline{\beta}}) = \sigma^2 (X'X)^{-1} X' V X (X'X)^{-1} \neq \sigma^2 (X'X)^{-1}$.
Typically $E(MSE) \neq \sigma^2$.
Remedy: use GLS if V is known,
sandwich estimator

§9.4 13] Find outliers (cases far away from the bulk of the data) with response plots and residual plots.

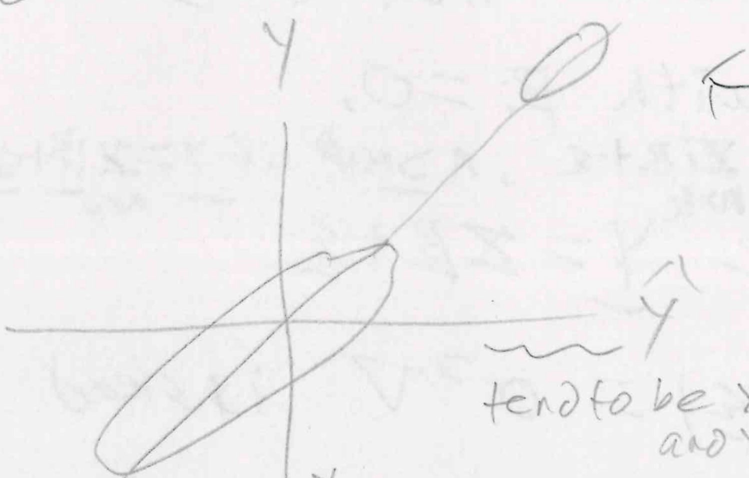


good response plot

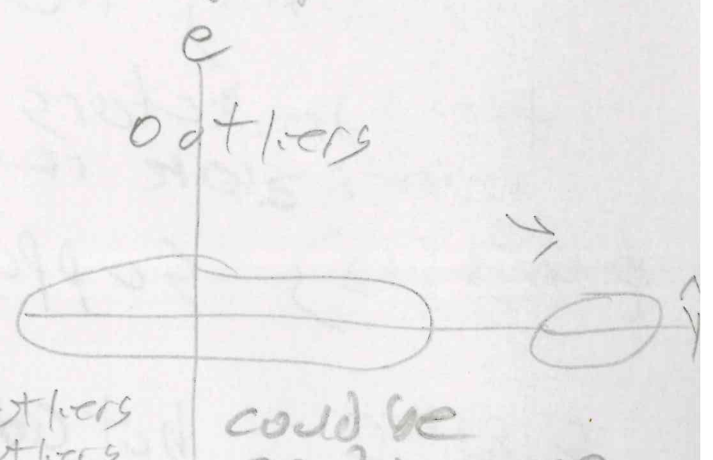


good residual plot

Outliers will often have gaps



tend to be X outliers and Y outliers



could be good leverage points

← Y outlier far from bulk of Y's

I have an R mpack function rmreg2.

14] Robust estimators can also be used.