MULTIPLE REGRESSION II: DIAGNOSTICS

9.1 Model Adequacy - Partial Regression Plots

9.2 Outlying Y Observations

√ residuals	e
√ semistudentized residuals	$e^* = e / RMSE$
new studentizedresiduals	$r = e / \{(1-h)^{1/2} RMSE\}$
new studentized deleted residuals	$t_i = d_i / RMSE_{[-i]}\{1-h)^{1/2}$

9.3 Outlying X observations

leverage

9.4 Influential cases

DF Fits

DF betas

Cooks Distance (aggregate effect) on

9.5 Collinearity Diagnostics - Variance Inflation Factors

9.1 Model Adequacy

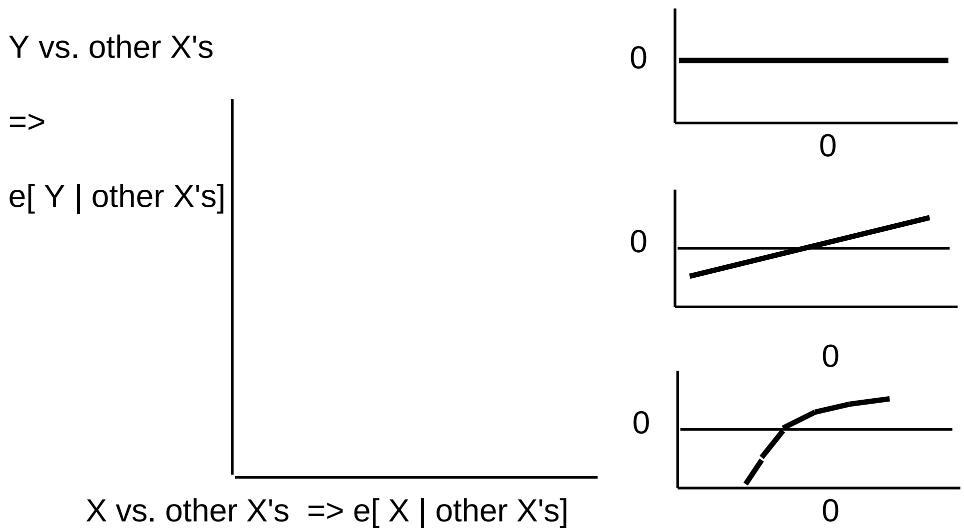
Plotting e(Y|X1,X2) vs. X2 may be misleading / inaccurate

- it ignores fact that X2 is already (partially) used via X1.

PARTIAL REG^n. PLOTS

also called "ADDED VARIABLE" or "ADJUSTED VARIABLE" PLOTS

• Marginal relation for X_k given other X's already in model



RECALL:

beta_k = slope of e(Y | other X's) on e[X_k | other X's]

2 Examples

1. Wrong impression if plot e(Y|X1,X2) vs X2 and vs X1

   - ok if plot e(Y| X2) vs e(X1|X2)

     e(Y| X1) vs e(X2|X1) *(Fig 9.3)*

   - *also note outlier in " e(x1/x2)" space*

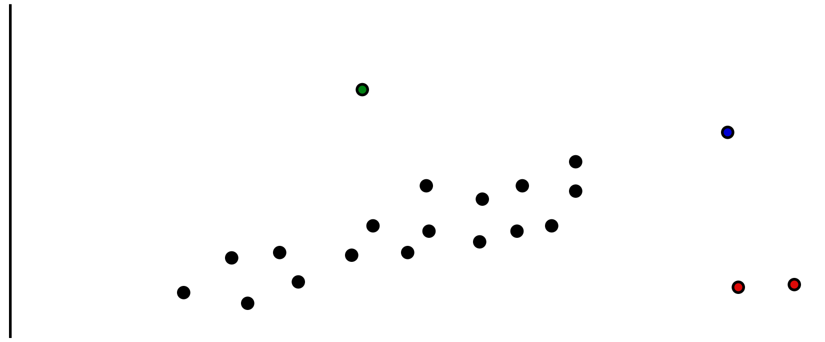
2. one of the 2 X's matters; other doesn't

   - wouldn't know just from e(Y|X1, X2) vs X1 or vs X2 *(Fig 9.4)*

   (high r between X1and X2)

9.2 Outlying Y Observations *Studentized* Residuals and *Studentized Deleted* Residuals

Observation can be outlying **in Y** or X or both (Fig 9.5)



Y outliers

residual: $e = Y - \hat{Y}$

Semi-studentized residual: $e^* = e / \text{RMSE}$

Why e* not always enough...

Residuals (e's) may have substantially different variances

Should scale each e by its own SD, rather than by just one $\hat{SD}$ = RMSE for all e's

How So?

$$\underset{n \times 1}{\hat{\mathbf{Y}}} = \underset{n \times n}{X(X^TX)^{-1}} \; \underset{n \times 1}{X^TY} = \underset{n \times 1}{H\,\mathbf{Y}}$$

H : "hat" matrix
see e.g. Table 9.2

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - H\mathbf{Y}$$

$$= \underset{n \times n}{(I - H)}\, \mathbf{Y}$$

$\text{Var}(\mathbf{e}) = \sigma^2 (I - H)$

$\text{Var}(e_i) = \sigma^2 (1 - h_{ii})$ Covar $(e_i, e_j) = \sigma^2 (1 - h_{ij})$

$\hat{\text{V}}\text{ar}(e_i) = \text{MSE}(1 - h_{ii})$

Why is var (e) $= \sigma^2 (1 - h_{ii})$?

- the more "outlying" the X, the more it affects e (makes it <u>smaller</u> on ave.)

NOTE: Notice difference between unobservable $e \sim N(0, \sigma^2)$ and calculable $e = Y - \hat{Y}$, whose variability is determined by design matrix (X) and where the X (associated with Y) is relative to other individual X's. (e.g. "3 point" regression)

Refined Residuals

(1) First Refinement Scale each e by SD(e) i.e.

$r_i = e_i / \{ \text{RMSE} (1 - h_{ii})^{1/2} \}$

Then r's have same variance

("INTERNALLY STUDENTIZED"... MSE from entire dataset incl. Y_i)

(2) Second Refinement (TO BE MORE SENSITIVE)

If Y_i an outlier, it will make (R)MSE artificially large and thus r_i artificially small.

Solution: : exclude Y_i from calculation of MSE

1) use $d_i = y_i - \hat{Y}_{[-i]}$ where $\hat{Y}_{[-i]}$ is calculated form other n-1 observations.

2) use MSE from dataset that excludes Y_i

$t_i = d_i / (\text{var}[d_i])^{1/2}$ ***(" Studentized deleted residuals ")***

Computationally: $d_i = e_i / (1 - h_{ii})$ (" h_{ii} " influence)

$\text{Var}(d_i) = \sigma^2{}_{[-i]} / (1 - h_{ii})$

$$t_i = d_i / (\text{var}[d_i])^{1/2} = \frac{e_i / (1 - h_{ii})}{\sqrt{\text{MSE}[-i] / (1 - h_{ii})}}$$

$$= \frac{e_i}{\sqrt{\text{MSE}[-i] / (1 - h_{ii})}}$$ *"Externally studentized"*

$t_i \sim t_{n-1-p}$ ...but not independent

How to obtain $\text{MSE}_{[-i]}$ without n separate "refits"

$t_i = e_i \sqrt{\dfrac{n - 1 - p}{\text{SSE}(1-h_{ii}) - e_i^2}}$ So just need e's SSE h_{ii} <-- key.

Largest $|t_i|$ ...n tests (Bonferroni)