Some traditional diagnostics

- residual analysis
- case analysis

Plots of Residuals

- Plot $\hat{\boldsymbol{\varepsilon}}$ vs. $\hat{\mathbf{Y}}$
  – assess for heteroscedasticity
  – assess for trends indicating poor model fit
  idealized examples on overhead

  NOTE: Residual plots can exaggerate trends. This can be a good thing, but scale is important.

  ISSUE: Using the data to develop the model can lead to overfitting and over-confident inference.

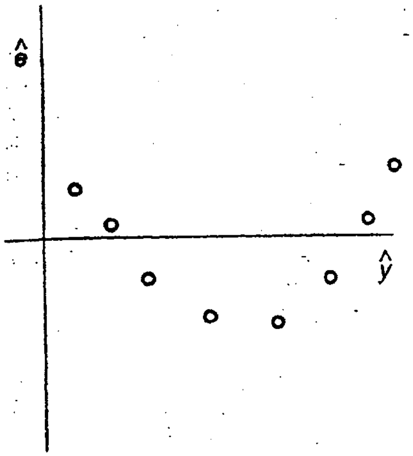
- Normal probability plot or QQ plot
  –assess for deviations from normality
  idealized examples on overhead

  ISSUE: The normality assumption is most important with small samples. Unfortunately, this is when QQ plots are the most difficult to interpret.

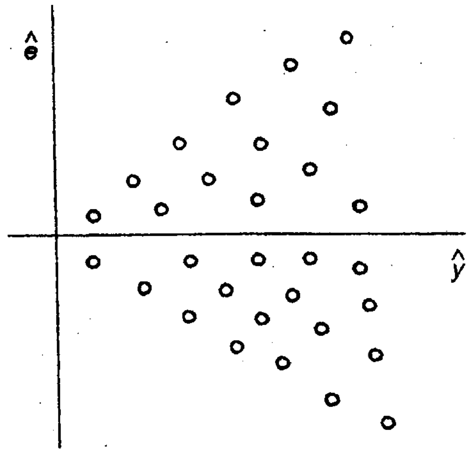
  EXAMPLE: The sets of figures A, B, C, show six samples of size 25 from a normal distribution, double-exponential distribution (heavier tails), and Cauchy distribution (heaviest tails). Which is which?

  A=normal; B-double exponential; C=Cauchy

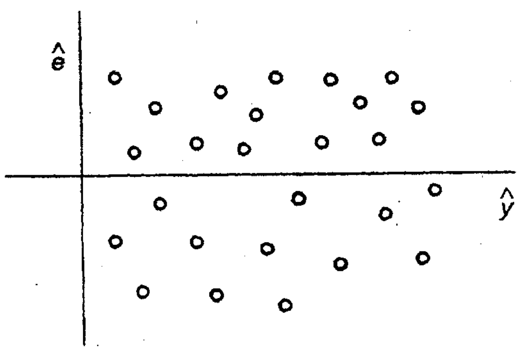
NOTE: Both of the diagnostics above work regardless of the number of predictors.



(a) Biased errors

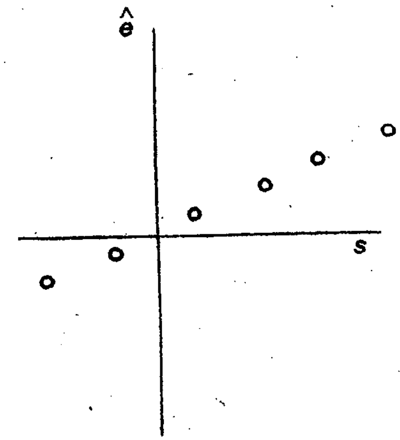


(b) Non-constant variance

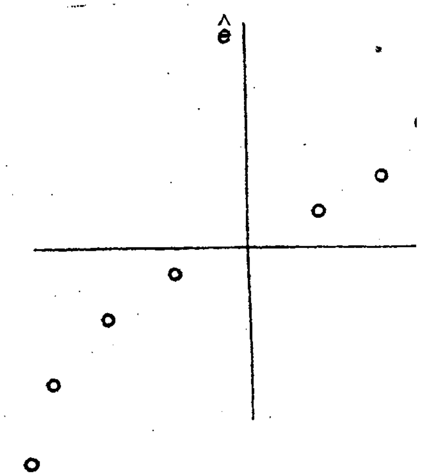


(c) Looks good

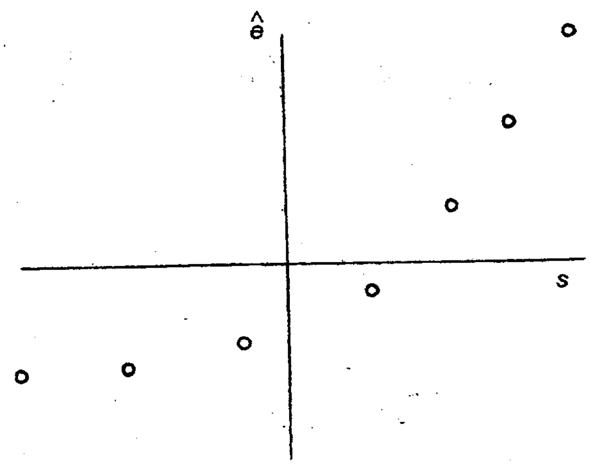
$\hat{\varepsilon}$ vs. $\hat{Y}$



(a) Normal



(b) Long or high-tailed



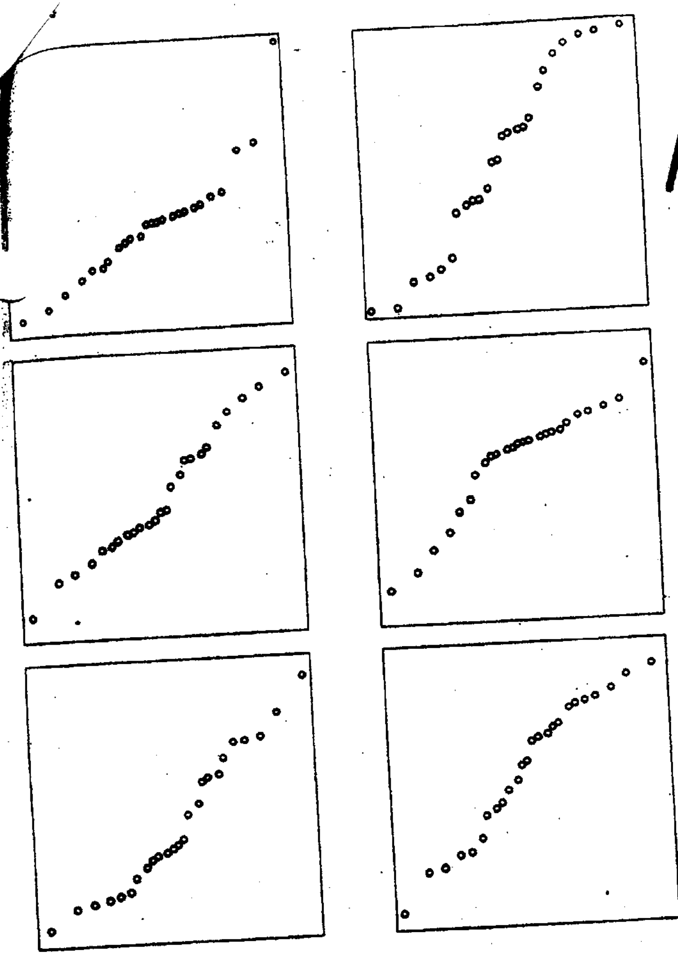
(c) Skewed right

Normal probability plots.

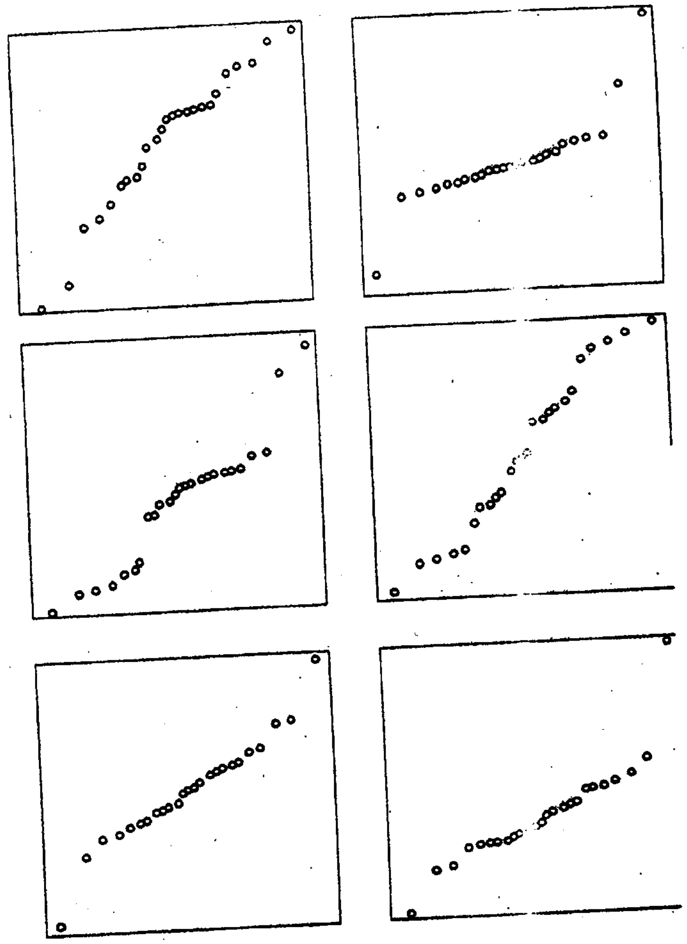
QQ Plots

1

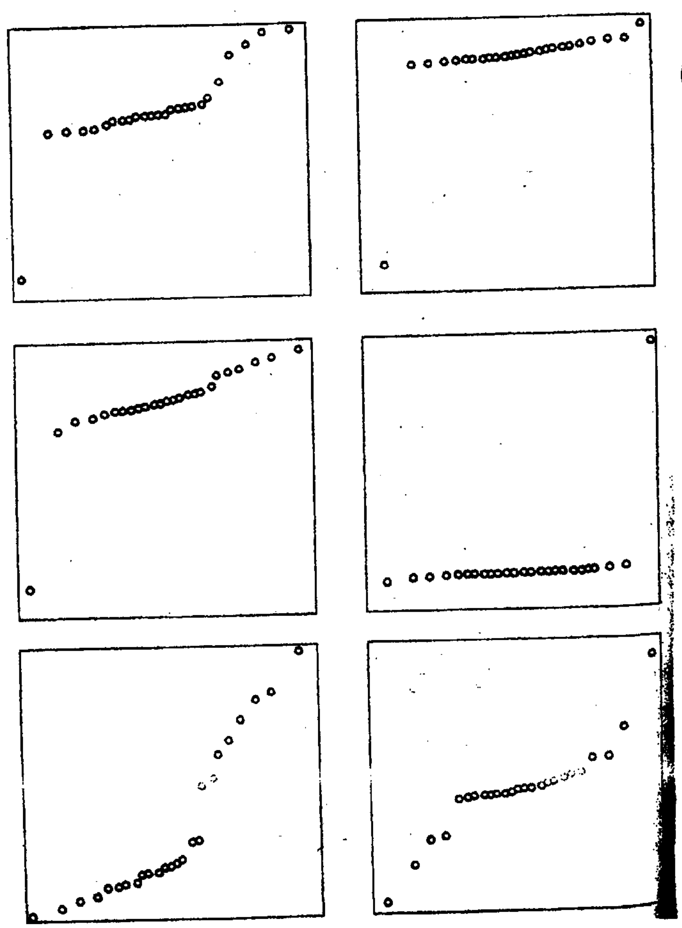
A



B



C



Fun with the Hat Matrix

$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
(the matrix formerly known as $\mathbf{P}$).

We know: $\mathbf{H} = \mathbf{H}' = \mathbf{H}^2$

We know: $\mathbf{H}\hat{\boldsymbol{\varepsilon}} = 0$, i.e. $\sum_j h_{ij}\hat{\epsilon}_j = 0 = \sum_j h_{ji}\hat{\epsilon}_j$

Since $\mathbf{H}^2 = \mathbf{H} = \mathbf{H}\mathbf{H}$,
$h_i \equiv h_{ii} = \sum_j h_{ij}h_{ji} = \sum_j h_{ij}^2 \geq 0$

We know: $\text{trace}(\mathbf{H}) = \text{rank}(\mathbf{H}) = p = \text{rank}(\mathbf{X})$ so $\sum h_i = p$ and since $h_i \geq 0$ the "average" value of h_i is p/n.

Cases i with $h_i \geq 2p/n$ are sometimes labeled "high leverage" cases (to be explained).

Writing

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix},$$

then we can write $h_{ij} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_j$.

We may want to identify:

Leverage cases - remote in the predictor space. May or may not influence the parameter estimates but may influence other summaries such as R^2 or $\hat{\sigma}^2$.

Influence cases - have a noticeable effect on the estimated regression coefficients.

If all we ever did was simple linear regression we might not need special ways to identify leverage cases or influence cases. However, these cases can be difficult to identify with higher-dimensional models.

idealized examples on overhead

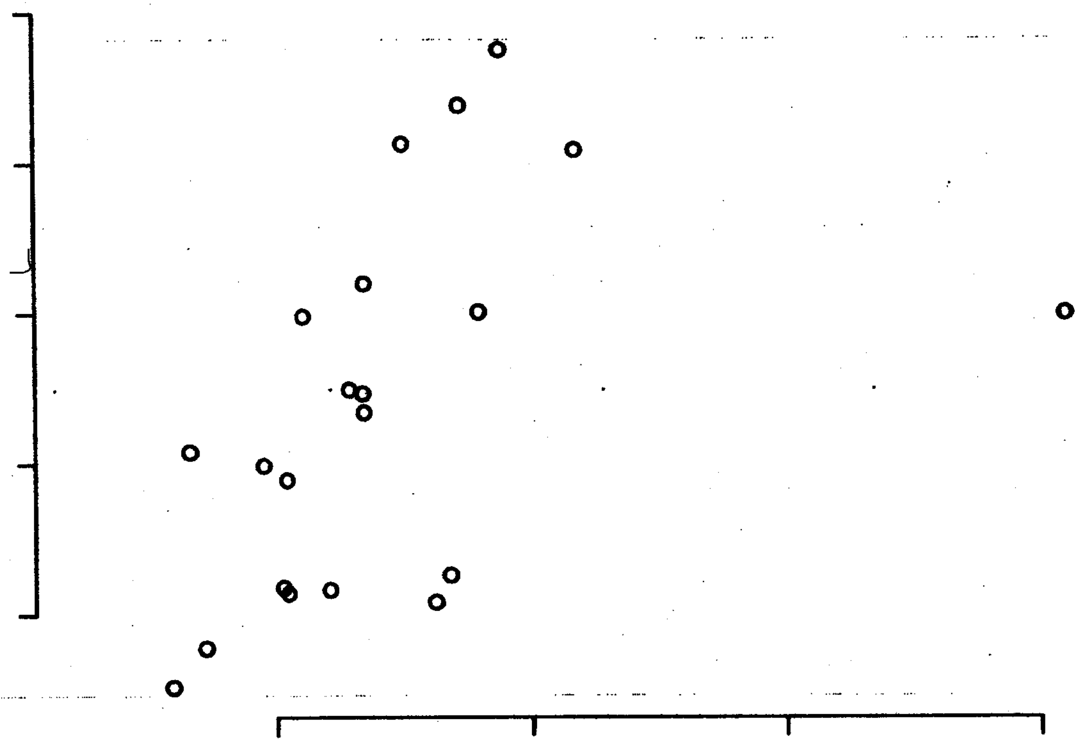
NEXT:

- h_i is a measure of leverage
- For influence, we want to compare $\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}(i)$, the parameter estimates from fitting the model without case i.
  $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}(i)$ are related via $\mathbf{H}$.

ALSO:

- Studentized residuals, a better residual for diagnostics.
- RSS and $RSS(i)$ are also related via $\mathbf{H}$.

Influential Observation



High leverage observation