Model Checking and Refinement

- Graphical Analysis (scatter plots)
- Fit Regression and Obtain Residual Plots
- Transformations ?
- Outliers or Influential Cases? Investigate:
  - Leverage
  - Cook's Distance
  - Studentized Residuals
- Refine Model by removing variables

Leverage

- Leverage of a case is a measure of the distance between its explanatory variable values and the average of the explanatory variable values
- With one explanatory variable:

$$h_i = \frac{1}{n} + \frac{1}{n-1}\left[\frac{X_i - \bar{X}}{S_X}\right]^2$$

- In general, can compute leverage as

$$h_i = \left[\frac{SE(fit_i)}{\hat{\sigma}}\right]^2$$

Leverage...

- High leverage implies case has a high potential for influence
- Case occupies a position in the X-space that is not densely populated by other cases, so this case can draw the regression toward it
- leverage is between $1/n$ and 1
- the average leverage is p/n
- depends on the X's in the model

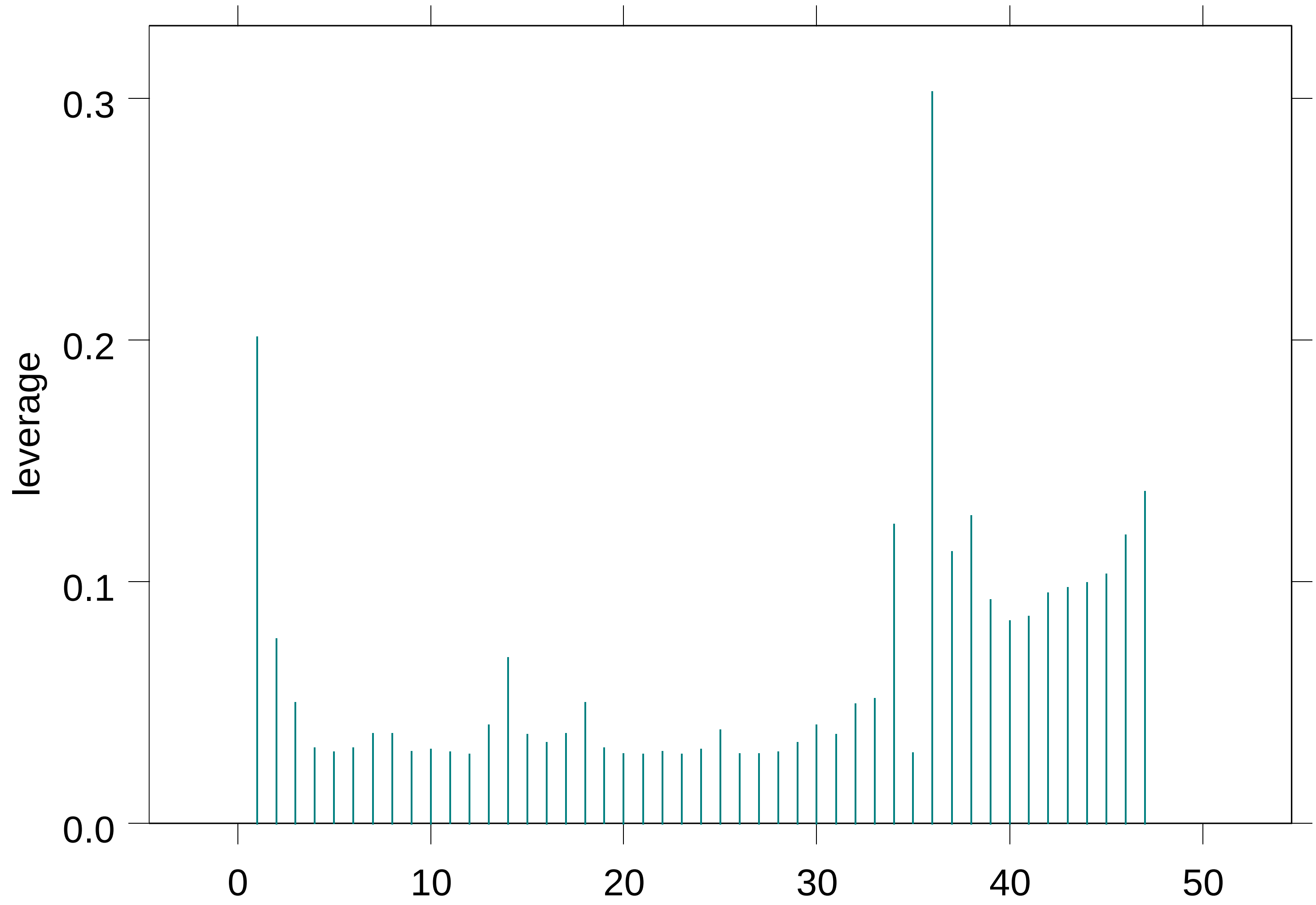
Leverage...

- A case i with a large leverage has a residual with low variability:

$$SD(residual_i) = \sigma \sqrt{1 - h_i}$$

- implies residual must be small and draws regression to it
- if the regression line based on the other points goes close to the case, then it is not necessarily influential

Leverage Plot for Bee Example



Cook's Distance

- Measure of overall influence
- Effect of dropping case i on the fitted values

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_{j(i)} - \hat{Y}_j)^2}{p\,\hat{\sigma}^2}$$

$$D_i = \frac{1}{p}\left(\frac{residual_i}{\sigma\sqrt{1-h_i}}\right)^2 \left(\frac{h_i}{1-h_i}\right)$$

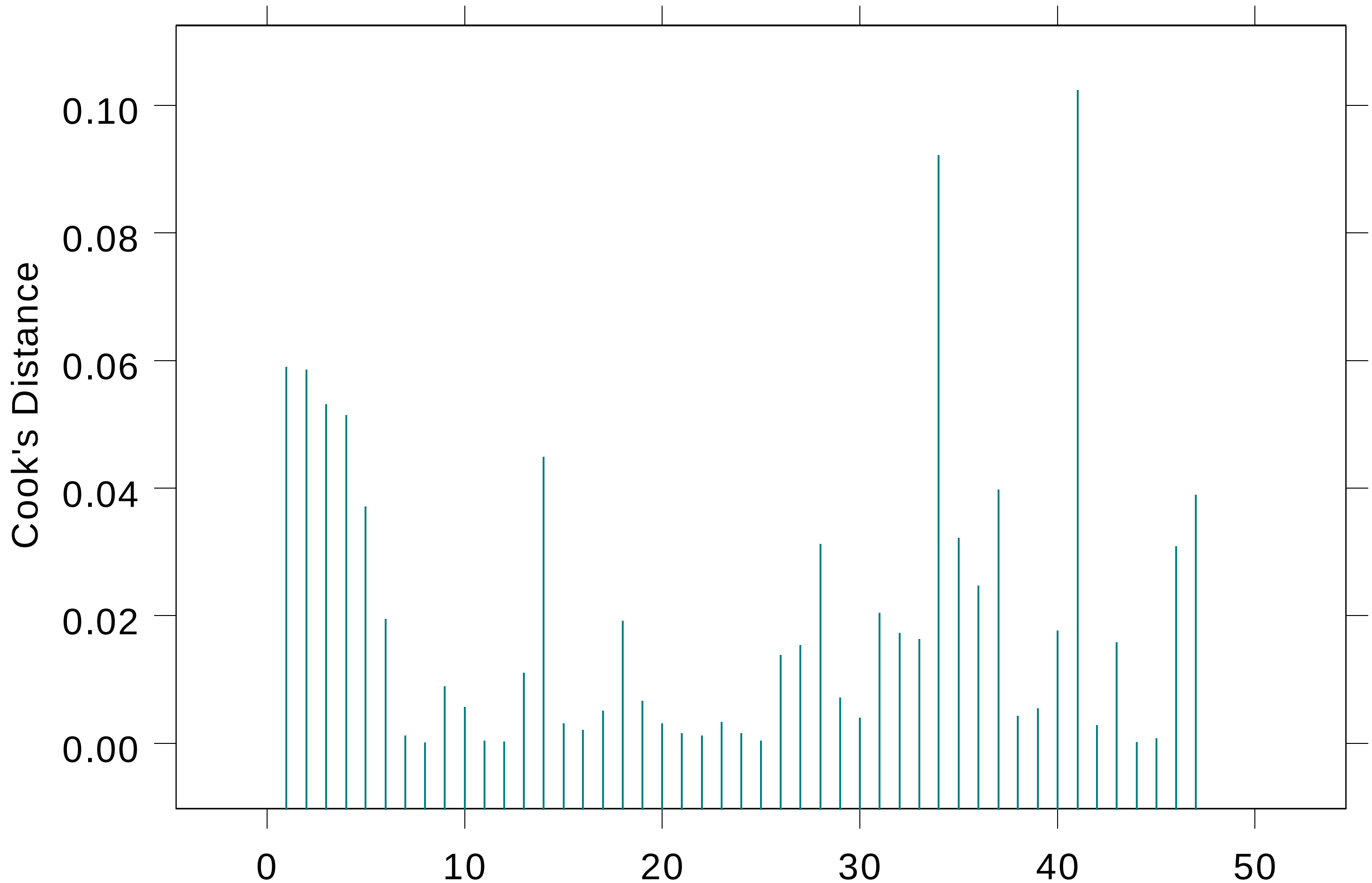
Internally Studentized residual

or standardized residual

Cook's Distance

- Cook's distance of a case is big (influential) if
  - h_i leverage is large
  - standardized residual is large
  - or both
- Cook's Distance > 1 should be examined!

Cook's Distance for the Bee Ex



Externally Studentized Residuals

- Standardized or internally studentized residuals use an estimate of σ based on all the data
- not robust if there is an outlier
- Externally Studentized residuals omit the ith case when estimating σ and standardizing the residual

$$studres_i = \frac{residual_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

Externally Studentized Residuals

- Values of |studres| > 2 should be investigated as potential outliers
- Model for the ith observation being an outlier:

$$\mu_i = \beta_0 + \beta_1 \log(duration_i) + \beta_2 I(code)_i + \delta I_i + \epsilon_i$$

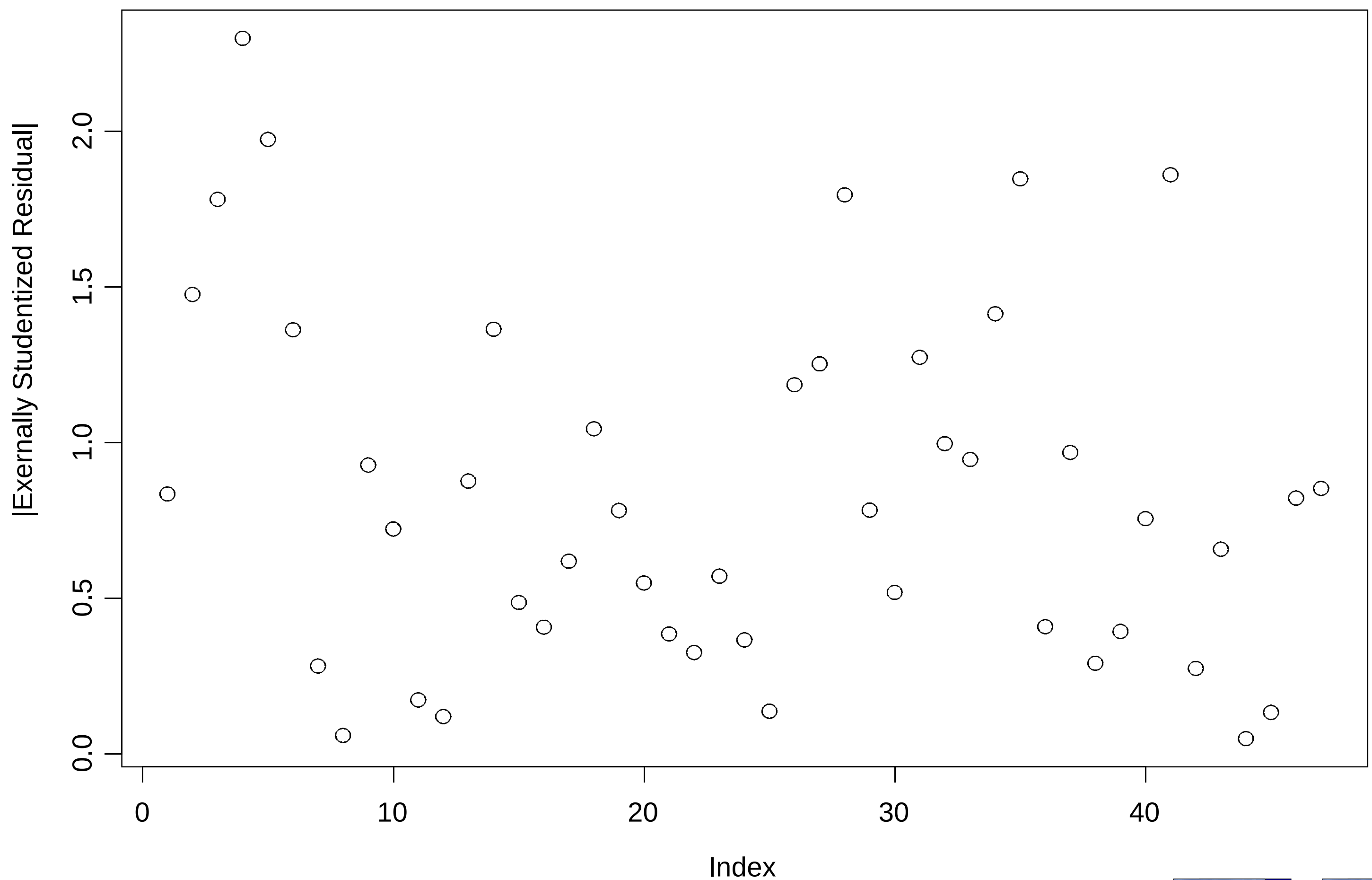
- I_i is 1 for the ith oservation and 0 otherwise
- Fits a separate mean δ for the ith case
- Test Ho: $\delta = 0$

Outlier Test

- The externally studentized residual is the t-statistic for testing Ho: $d = 0$ for the ith case
- For informal checking, check $|studres| > 2$
- To avoid problems with multiple hypothesis testing, use the Bonferroni inequality
- Compare the p-value to α/n
- p-value is based on $(n - p - 1) - 1$ df

(n - total # parameters in the mean)

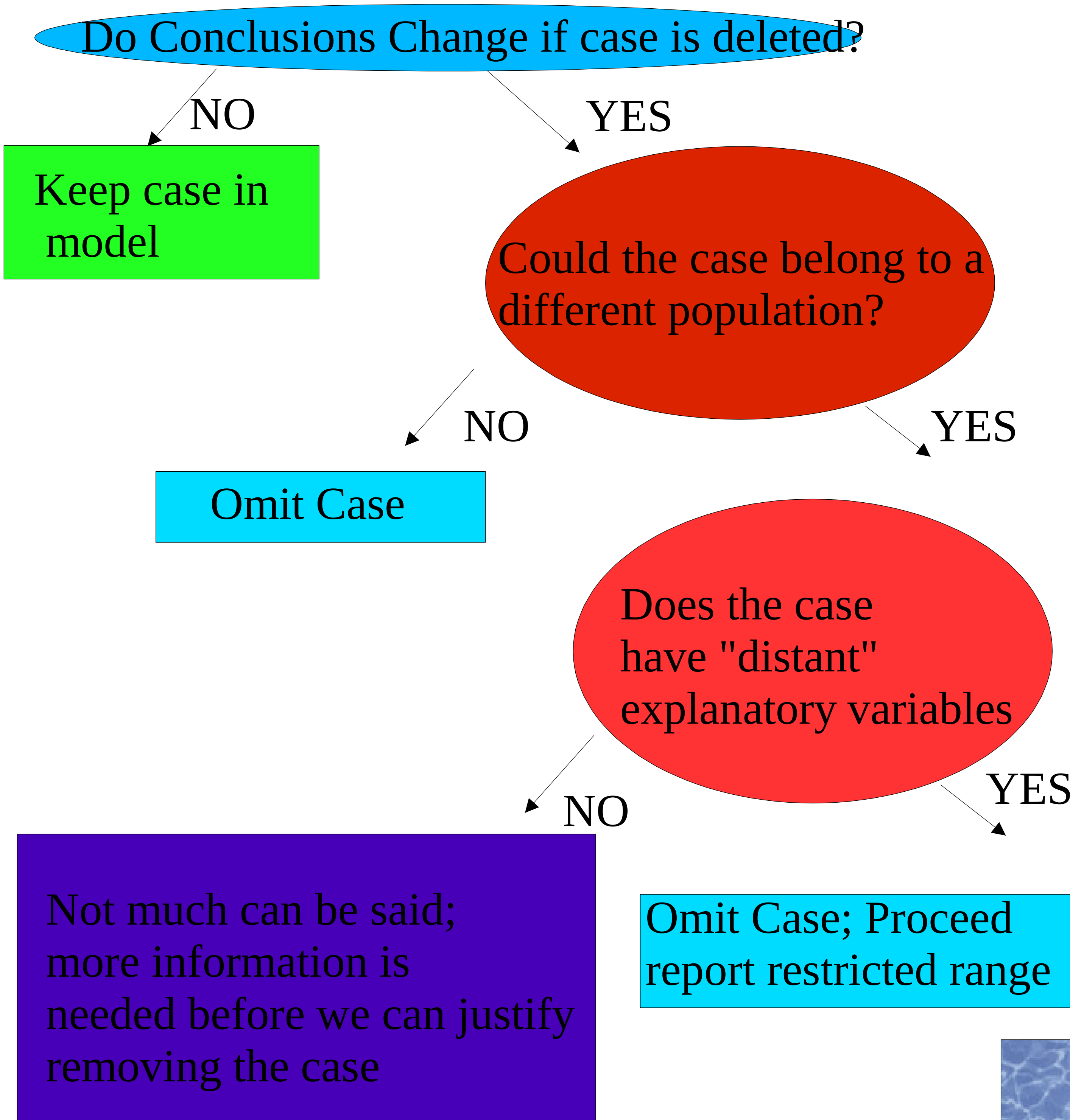
Absolute value of Externally Studentized Residuals



Summary of Diagnostics for the Bees Example

Case Diagnostics

- Maximum leverage = 0.30; case 36
  - not influential
- maximum Cook's Distance = 0.10; case 41
  - not influential
- max |externally studentized residual| = 2.299
  - p-value = 0.026; compare to .05/47 = 0.001
  - Do not reject Ho: $\delta = 0$; case 4 is not an outlier
- Everything checks out!


Do Conclusions Change if case is deleted?
NO
YES
Keep case in model
Could the case belong to a different population?
NO
YES
Omit Case
Does the case have "distant" explanatory variables
NO
YES
Not much can be said; more information is needed before we can justify removing the case
Omit Case; Proceed report restricted range