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

Leverage...
- High leverage implies case has a high potential for influence on the regression line
- 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 \(\frac{1}{n}\) and 1
- the average leverage is \(\frac{p}{n}\), \(p=\#\text{parameters}\)
- depends on the X's in the model

Leverage...
- A case \(i\) with a large leverage has a residual with low variability:
  \[ SD(\text{residual}_i) = \sigma \sqrt{1-h_i} \]
- 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

\[ D_i = \frac{\sum (\hat{Y}_{j(i)} - \hat{Y}_i)^2}{p \sigma^2} \]

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 residual*: how far is $y_i$ from $\hat{y}_i$ in terms of SE's.
- Standardized (or internally studentized) residuals use an estimate of $\sigma$ based on all the data.
- not robust if there is an outlier
- Externally Studentized residuals omit the $i$th case when estimating $\sigma$ and standardizing the residual

$$studres_i = \frac{\text{residual}_i}{\hat{\sigma}_i \sqrt{1-h_i}}$$

**Outlier Test**

- The externally studentized residual is the test statistic for testing $H_0: \delta = 0$ for the $i$th case
  - For informal checking, check $|studres_i*| > 2$
  - Under $H_0$, $studres_i* \sim t$ with d.f.=$n$-$\#$parameters
    - Here, there are $n$-$4$ parameters (including $\delta$)
  - $\max|t_i|$ indicates a significant outlier at level $\alpha$ if $t_i$ is significant at level $\alpha/n$.
  - Can calculate the p-value and compare to $\alpha/n$

**Externally Studentized Residuals**

- Values of $|studres_i*| > 2$ should be investigated as potential outliers.
- Model for the $i$th 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 $i$th observation and 0 otherwise
- Fits a separate mean $\delta$ for the $i$th case
- Test $H_0$: $\delta = 0$

**Absolute value of Externally Studentized Residuals**

![Graph showing the 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: δ = 0; case 4 is not an outlier

Do Conclusions Change if case is deleted?

- NO
  - Keep case in model

Could the case belong to a different population?

- NO
  - Omit Case
- YES
  - Does the case have "distant" explanatory variables?
    - NO
      - Omit Case
    - YES
      - Not much can be said; more information is needed before we can justify removing the case

Omit Case; Proceed report restricted range