DIAGNOSTICS

**Questions:**

- Are model assumptions satisfied?
- Is the result unduly influenced by one or a small number of points?

We've discussed some techniques to study these questions:

- Normal plots
- Scatterplots
- Lowess, lowess±SD
- Remove linear trend
- Leverage – identify some potentially influential points

More techniques:

**Caution:** Not guaranteed to catch all problems, but can catch some.

---

1. **RESIDUAL PLOTS**

**Recall the error formulation of the OLS model:**

Ordinary Linear:

\[ \mathbf{Y|X} = \mathbf{\eta}^\top \mathbf{u} + \epsilon \mathbf{X} \]
\[ \epsilon \mathbf{X} \sim \mathcal{N}(0, \sigma^2), \text{ independent of } \mathbf{X} \]

**Recall the Least Squares fits and residuals**

OLS:

\[ \hat{y}_i = \hat{\eta}^\top \mathbf{u} \]
\[ \hat{e}_i = y_i - \hat{y}_i \]

Intuitively, \( \hat{e}_i \) is an estimate of \( e_i \) ( = \( y_i - \hat{\eta}^\top \mathbf{u} \)).

We know \( \text{E}(e_i) \) and \( \text{E}(\hat{e}_i) \) are both zero, so \( \hat{e}_i \) is an unbiased estimate of \( \text{E}(e_i) \).

Thus it seems reasonable to plot the \( \hat{e}_i \)'s against various things to give some check on whether the model assumptions are reasonable.
However:

- \( \text{Var}(e_i | x_i) = \sigma^2 \)
- \( \text{Var}(\hat{e}_i | x_i) = \sigma^2 (1 - h_i) \), where \( h_i \) is the \( i \)th leverage.  
  (Section 7.6 of book)

Thus:

- If the \( h_i \)’s are all small, then
  \[ \text{Var}(\hat{e}_i | x_i) = \text{Var}(e_i | x_i), \]
  so a plot using the \( \hat{e}_i \)'s should approximate a plot using the \( e_i \)'s.

- If the \( h_i \)’s are all approximately equal, then
  \[ \text{Var}(\hat{e}_i | x_i) \text{ is approximately a constant times } \text{Var}(e_i | x_i), \]
  so a plot using the \( \hat{e}_i \)'s should approximate a rescaled plot using the \( e_i \)'s.

- But if the \( h_i \)'s vary noticeably, then a plot using the \( \hat{e}_i \)'s will not give a good approximation of a plot using the \( e_i \)'s. In this case, use instead
  \[ \hat{e}_i = \frac{\hat{c}_i}{\hat{\sigma} \sqrt{1 - h_i}}. \]
  (These are automatic in some software; not in arc)

- Studentized residuals should have something near a standard normal distribution, so are helpful in identifying extreme cases.

To form these in arc:

i. Select “Add to data set” from model menu.  
   Select L1:Residuals and L2:Leverages

ii. Note that the variables L1:Residuals and L2:Leverages are added (Punctuation important!)

iii. Use “Add a variate” on the data menu to define a new variable \( sr \) in terms of the newly added variables and the value of \( \hat{\sigma} \) from the regression output.
Types of Residual Plots (roughly in order of importance)

- Against fitted values $\hat{y}_i$
- Against individual or pairs of predictors
- Against other possible predictors not in the model (especially time, location)
- Against individual terms other than predictors
- Against linear combinations of terms

Suggestions for Residual Plots for Specific Purposes

Checking linearity

Plot against fitted values $\hat{y}_i$. (Like "remove linear trend")

Checking constant variance

Plot against fitted values, predictors, pairs of predictors, other possible predictors.

Caution: What looks like non-constant variance can sometimes be caused by non-linearity.

Example: caution.lsp

Checking independence

Plot against other possible predictors (especially time, location)
Checking normality

Use a normal probability plot – using studentized residuals if warranted.

Checking for outliers:

Plot against fitted values, predictors, pairs of predictors
A “multipanel plot” can be useful.

II. COOK’S DISTANCE

Recall: An observation is influential if it has more of an effect on the OLS estimates than the other cases do.

With 1 or 2 terms, it's relatively easy to spot a potential influential point on the scatterplot and check if the point is influential. Leverage can help pick out x-outliers, which are potentially influential. Cook's distance can help check for influence more generally.

The idea:

• Delete the $i^{th}$ case and compute the least squares estimator without using the $i^{th}$ case.

• Evaluate this estimator at $\hat{x}_i$, giving $\hat{y}_{i,j}$

  $$\hat{y}_{i,j} = \text{the fit at } \hat{x}_i, \text{ not using the } i^{th} \text{ case}$$

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

measures the total influence of the $i^{th}$ case. (This can be expressed in terms of the coefficient estimates -- see more details in Section 15.2)
Rules of thumb in using $D_i$:

- Plot $D_i$ vs case number.

- Examine cases that have relatively large $D_i$ (i.e., large relative to other values for these data)

- Examine cases with $D_i > 0.5$, and especially cases with $D_i > 1$.

- There is no hypothesis test using $D_i$. 