Recap: Scatterplot and residuals plots may detect a violation of the model assumptions.

- (Studentized) Residuals vs. fitted values check Linearity and Equal variance assumptions
  - Focus on whether there is a pattern indicating curvature in original relationship
  - Focus on whether there is a band of constant width across the x values (e.g., no “fanning”)
- Histogram and normality plot of residuals checks for Normality assumption
  - Look for non-linear pattern in probability plot, not just outliers
  - Can “descriptively” use a goodness-of-fit test p-value
  - Can be less critical with larger sample sizes (except for prediction intervals)
- Later will look more carefully at Independence assumption

Often we can transform one or both variables to “fix” one or more of these issues. They can also help with data exploration regardless of trying to fit regression model assumptions. The question becomes how to efficiently select a useful transformation and then how to interpret the resulting relationship…

If the scatterplot is monotonic, nonlinear, and simple, we can often use a “power family transformation”

- These transformations can be applied to either X or Y or both, e.g., \( x' = x^{power} \)
- The degree of the transformation can be selected from the ladder: \(-1/y^2, -1/y, \log(y), \sqrt{y}, y, y^2, \ldots\)
  - Usually consider powers between -2 and 3, often fine to use one of: -1, -1/2, 0, 1/3, 1/2, 1, 2, 3
  - Is possible to derive the value of the power that minimizes SSE (e.g., Box-Cox transformations), then usually rounded to something that makes practical sense.
  - The minus sign with negative powers maintains the ordering of the observations.
  - Can use \((x + k)^p\) where \(k\) ensures all values are positive first
  - Only effective when the ratio of the largest value to the smallest value is large (can shift first)

To select the transformation based on a monotonic scatterplot, use the following chart

<table>
<thead>
<tr>
<th>Increase y power</th>
<th>Decrease y power</th>
<th>Decrease x power</th>
<th>Increase x power</th>
</tr>
</thead>
</table>

- If linearity is the only issue, it is often better to transform \(x\) because easier to interpret and transforming \(y\) can mess up other conditions (NE).
- If normality and/or equal variance are the issue, transforming \(y\) is usually better (e.g., if residuals are positively skewed, lower the power of \(y\)). May need to transform both equally to maintain linearity.
- The Box-Cox procedure numerically estimates the “best” power transformation of the response variable, model is \(y^\lambda = X\beta + \epsilon\) and estimates \(\lambda\) (e.g., minimize SSE) and then round to something meaningful (\(\lambda = 0\) corresponds to log transformation).
Example 1: Reconsider the relationship between *prestige* and *income* for 102 Canadian occupations.
(a) Produce a scatterplot with a smoother. Fit a linear regression model and examine the residual plots. Is the relationship linear? Is the relationship monotonic? Is the relationship simple?

(b) Using the transformation ladder, suggest two possible next steps. Considering the residual plots, which of the two might you suggest?

(c) Compare the scatterplots using (a) $p = 1/2$ and (b) the log transformation on *income*. Which transformation would you recommend and why or what try next?

(d) Suggest another approach to analyzing these data.

Example 2: Data (1998) on *infant mortality* (infant deaths per 1000 live births) and *gross domestic product (GDP)* *per capita* (in US dollars) for 207 countries are in UN.txt.
(a) What is noteworthy about the scatterplot of mortality vs. GDP? Is the relationship nonlinear and monotonic and simple? What transformations are suggested by the transformation ladder for this form?

(b) Examine the scatterplot of $\log_{10}$ mortality vs. GDP. What would you suggest next?

(c) Examine the scatterplot of $\log_{10}$ mortality vs. $\log_{10}$ GDP. Does this reveal any new insights about the relationship/unusual observations? What is the least squares regression equation?

(d) To interpret the slope, how do we “increase $x$ by 1” here?

(e) So then the predicted change in $y$, $\log_{10}mortality_1 - \log_{10}mortality_2 = -.493$. Use (two) laws for logarithms to simplify this expression to be in terms of *mortality*.

Practice problem: Fit log(efficiency) vs. mpg and interpret the resulting slope.