36-401 Chapter 5: Fitting Curves (11/1/01)

Kim Sellers and Brian Junker

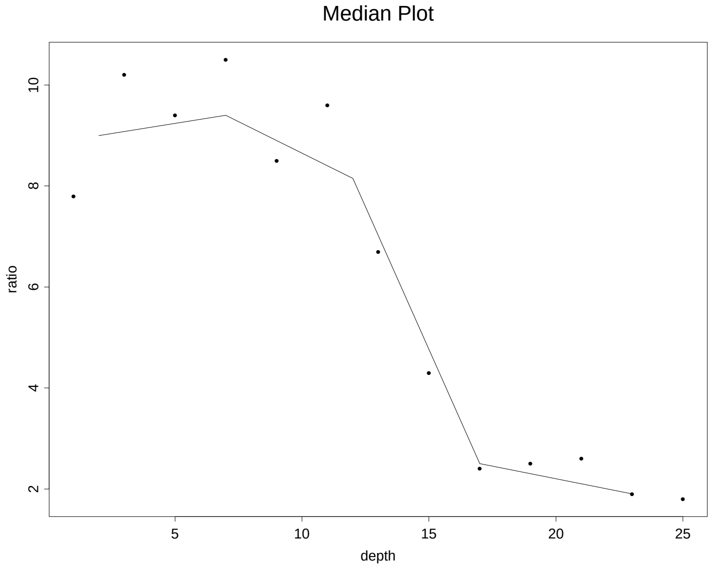
We experienced a bivariate form of a curvilinear regression in Chapter 2. In essence, we are interested in models that have an *intrinsically linear form*:

$$Y_i^* = \beta_0 + \beta_1 X_{i1}^* + \beta_2 X_{i2}^* + \ldots + \beta_{K-1} X_{i,K-1}^* + \epsilon_i,$$

where Y^* and $X_j^*, j = 1, \ldots, K-1$ are functions of Y and $X_j, j = 1, \ldots, K-1$ (e.g., $Y^* = Y^m$ for some m). Such a form allows us to use our current methodologies to solve curvilinear regression problems. However, models that cannot be expressed as shown above require nonlinear regression techniques.

Ways to fit curves to data:

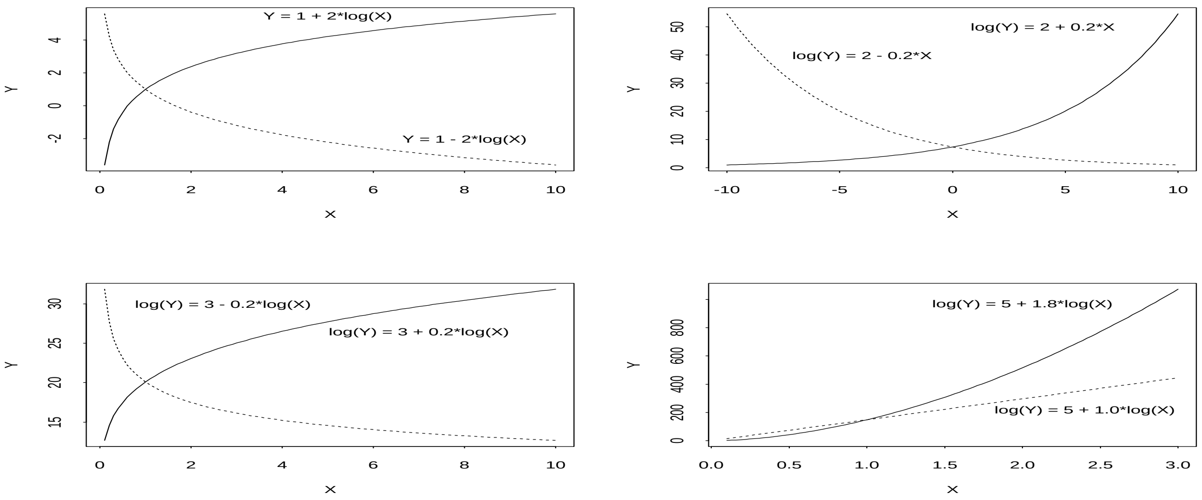
- Exploratory Band Regression: non-parametric guidance for parametric model fitting
  - Steps:
    1. Divide the scatterplot into a series of vertical bands
    2. Find central point (e.g., median) within each band
    3. Connect the resulting points to form curve
  - Useful to search for patterns in any two-variable scatterplot
  - Guide for parametric model fitting



- Regression with Transformed Variables
  - Transformations are considered to linearize $\mathbf{X}, Y$ relationship, sometimes at the risk of sacrificing normality, homoscedasticity, etc.

- Steps:
  1. Select appropriate transformations
  2. Generate transformed variables
  3. Using transformed variables, analyze data under new model
  4. After obtaining results regarding relationships, transform back to better interpret relations

- Curvilinear Regression Models
  - Example: Semilog curves and log-log curves have the forms $\hat{Y}_i = b_0 + b_1 \log X_i$, $\log \hat{Y}_i = b_0 + b_1 X_i$, or $\log \hat{Y}_i = b_0 + b_1 \log X_i$



  - The models $Y_i = \alpha e^{\beta X_i} \epsilon_i$ and $Y_i = \alpha X_i^{\beta} \epsilon_i$ can be transformed using logs to form intrinsically linear models.
  - Polynomial regression is a way to fit nonmonotonic curves
    * Second order: $\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2$

    * Third order: $\hat{Y}_i = b_0 + b_1 X_i + b_2 X_i^2 + b_3 X_i^3$

- - Higher order models may better fit the data, but are difficult to understand
  - Problems with polynomial regression:

    influence from outliers

    multicollinearity

    Thus, model is case dependent on its sample and may not effectively represent the big picture. Centering the X's about their mean $\bar{X}$ will reduce multicollinearity among powers of X but does not necessarily provide more precise estimates

- Conditional Effect Plots
  - Plots X_j against predicted Y with $X_i, i \neq j$ constant (substitute mean values for X_i's) to develop relationship between X_j and Y; note that a different set of choices for X_i's only affects the intercept
  - Comparing effects is easier for graphs containing the same Y scale; this allows us to compare effect strength for linear, curvilinear, or nonlinear models.

- Non-linear Regression
  - Arguments for nonlinear modeling:
    - Data suggest a specific nonlinear model
    - Nonlinear equation expresses our understanding of the mechanism causing changes in Y

  - Many nonlinear models to choose from. Below are a few examples:

    Non-linear exponential models

$$Y_i = \alpha e^{\beta X_i} + \epsilon_i$$

Negative exponential models (rise to asymptote with $\alpha, \beta > 0$)

$$Y_i = \alpha(1 - e^{-\beta X_i}) + \epsilon_i$$

Two term exponential models

$$Y_i = \frac{\theta_1}{\theta_1 - \theta_2}\left(e^{-\theta_2 X_i} - e^{-\theta_1 X_i}\right) + \epsilon_i$$

Logistic models (example of sigmoid or S-shaped curve)

$$Y_i = \frac{\alpha}{(1 + \gamma e^{-\beta X_i})} + \epsilon_i$$

Gompertz models (asymmetric S shape)

$$Y_i = \alpha e^{-\gamma e^{-\beta X_i}} + \epsilon_i$$

- Minimizing residual sum of squares requires iterative techniques. Unfortunately, methods aren't foolproof:
  * Need starting values.
  * May not converge.
  * May converge to local minimum instead of global minimum.

  Good initial estimates improve chances of an optimal nonlinear fit. Approaches to obtain initial parameter estimates include using comparable transformed linear form or estimating parameters from scatterplots
- R^2 defined as $1 - \frac{\text{RSS}}{\text{corrected TSS}}$. May be negative.
- Under normal i.i.d. errors, estimates and tests are *asymptotically* correct. Use the usual graphs to check model assumptions.