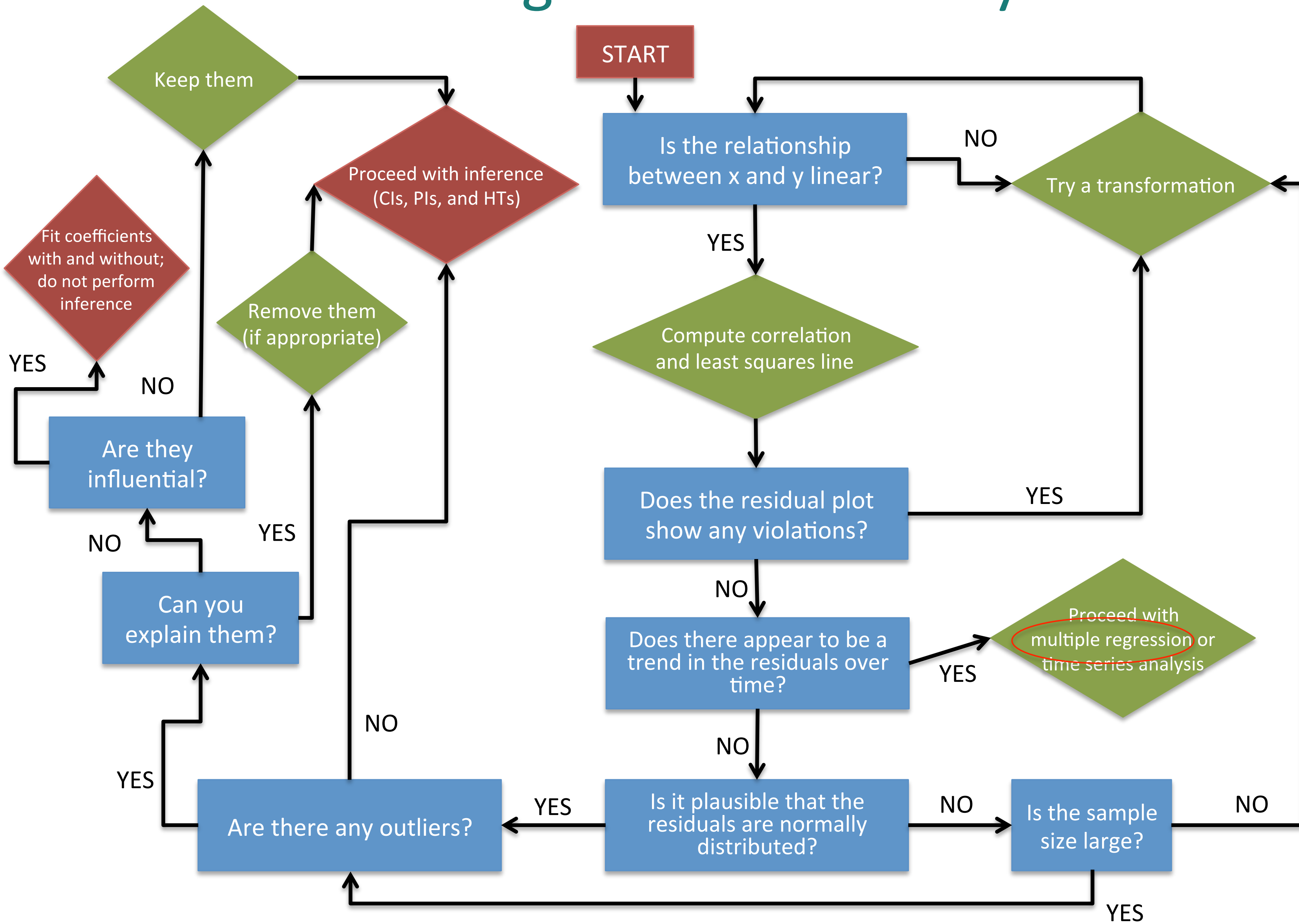
Introduction to Multiple Regression

Keegan Korthauer

Department of Statistics

UW Madison

SLR Diagnostics Summary



Multiple Linear Regression (MLR)

- SLR: models the relationship between a response variable (y) with a single predictor (x)

- When the response variable (y) actually depends on several factors (x_1, x_2,…, x_p) we can use a **multiple linear regression model**

- Many of the ideas and general concepts we learned for SLR are also applicable for MLR
  - Sums of squares
  - Assumptions 1 through 4 on the errors
  - Interpretation of coefficients
  - Diagnostics, etc…

Basic MLR Model

- Dependent continuous variable **y**
- *p* independent continuous variables $\mathbf{x_1, x_2, \ldots, x_p}$
- *n* observations: ordered pairs $\mathbf{(y_i, x_{1i}, x_{2i}, \ldots, x_{pi})}$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi} + \varepsilon_i$$

- Predicted y_i for a set of $x_{1i}, x_{2i}, \ldots, x_{pi}$:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \ldots + \hat{\beta}_p x_{pi}$$

Variations on the MLR Model

- Polynomial regression model
  - Dependent continuous variable y
  - *p* **degrees** of one independent variable x
  - *n* observations: ordered pairs $(\mathbf{y_i}, \mathbf{x_i})$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots + \beta_p x_i^p + \varepsilon_i$$

- Quadratic model of two independent variables
  - Dependent continuous variable y
  - 2 degrees of two independent variables x_1 and x_2
  - *n* observations: ordered pairs $(\mathbf{y_i}, \mathbf{x_{1i}}, \mathbf{x_{2i}})$

Interaction

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \beta_4 x_{1i}^2 + \beta_5 x_{2i}^2 + \varepsilon_i$$

Least Squares Coefficients

- Minimize the **sum of squared residuals (SSE)** to obtain coefficient point estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$
  - Analogous to SLR, involves taking p+1 partial derivatives, setting them equal to zero, and solving a system of p+1 equations…
  - OR a much more elegant expression using linear algebra…
  - But we'll rely on R to calculate the values for us

- SSE is still the sum of squared differences between observed y and predicted ŷ – no longer can visualize in 2D

$$e_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \ldots - \hat{\beta}_p x_{pi}$$

$$SSE = \sum_{i=1}^{n} e_i^2$$

Next

- More on Multiple Linear Regression
  - Interpreting Coefficients
  - Obtaining Estimates
  - Performing Inference
  - Diagnostics