Lecture 6

Simple Linear Regression V & Introduction to Multiple Linear Regression

Reading: Chapter 11, 12

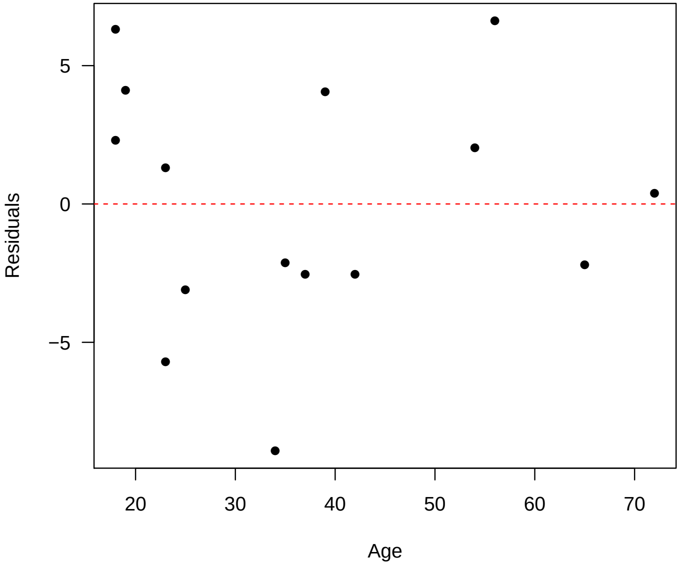
STAT 8020 Statistical Methods II
September 2, 2019

Whitney Huang
Clemson University

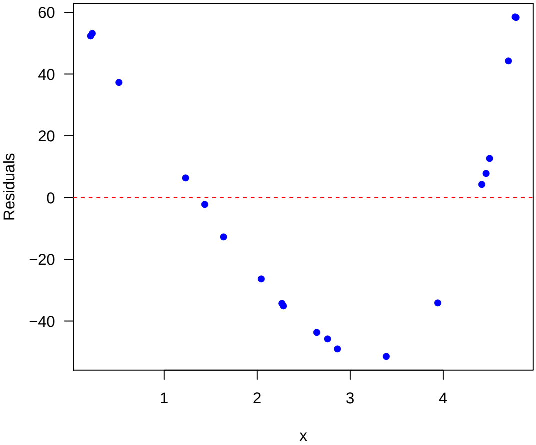
Agenda

1. **Regression Diagnostics and Remedies**
2. **Multiple Linear Regression**

MaxHeartRate vs. Age Residual Plot Revisited



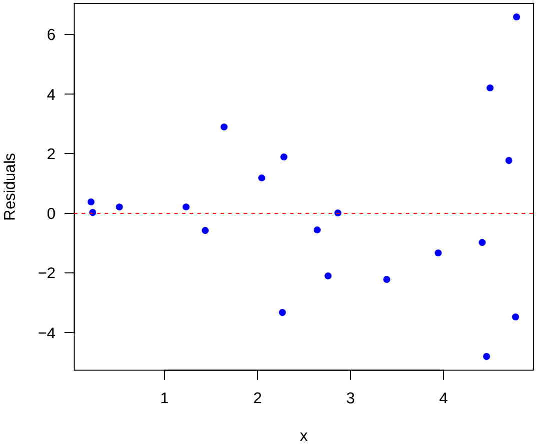
A Non-Linear Pattern



Possible Remedies:

- Transform X
- Nonlinear regression

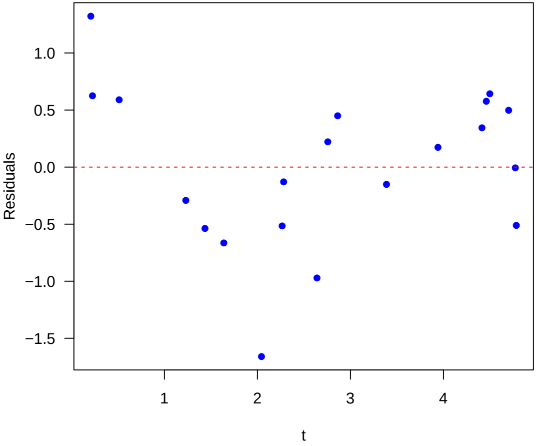
Non-Constant Variance



Possible Remedies:

- Transform Y
- Weighted least squares

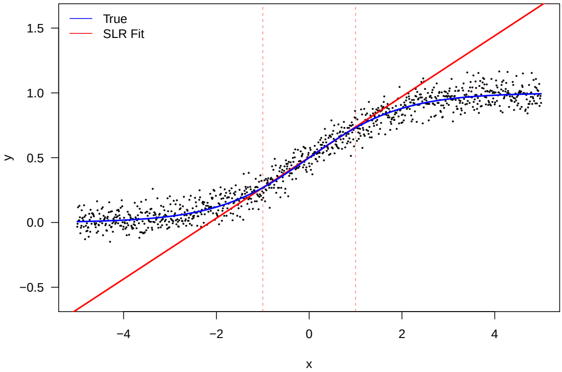
Correlated Errors



A Possible Remedy:

- Allow correlated errors in SLR

Extrapolation in SLR



Extrapolation beyond the range of the given data can lead to seriously biased estimates if the **assumed relationship does not hold the region of extrapolation**

Summary of SLR

- **Model:** $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- **Estimation:** Use the method of least squares to estimate the parameters
- **Inference**
  - **Hypothesis Testing**
  - **Confidence/prediction Intervals**
  - **ANOVA**
- **Model Diagnostics and Remedies**

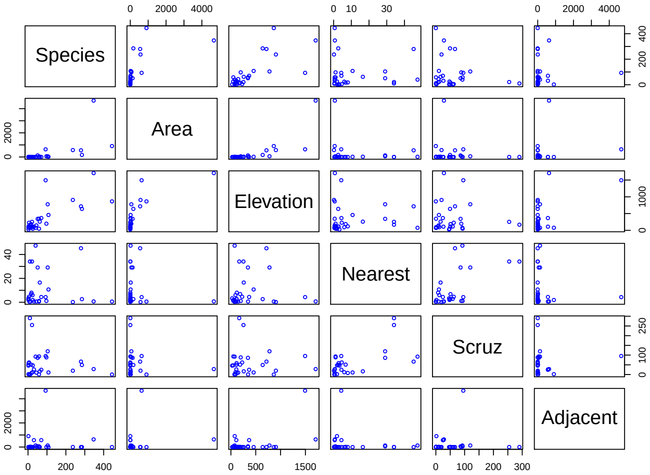
Multiple Linear Regression

Goal: To model the relationship between two or more explanatory variables (X's) and a response variable (Y) by fitting a **linear equation** to observed data:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$

Example: Species diversity on the Galapagos Islands. We are interested in studying the relationship between the number of plant species (`Species`) and the following geographic variables: `Area, Elevation, Nearest, Scruz, Adjacent`.

How Do Geographic Variables Affect Species Diversity?



$$\texttt{Species} = \beta_0 + \beta_1\texttt{Area} + \beta_2\texttt{Elevation} + \beta_3\texttt{Nearest} + \beta_4\texttt{Scruz} + \beta_5\texttt{Adjacent} + \text{error}$$

Fit a Multiple Linear Regression using R

```
lm(formula = Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
    data = gala)

Residuals:
     Min       1Q   Median       3Q      Max
-111.679  -34.898   -7.862   33.460  182.584

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06
Nearest      0.009144   1.054136   0.009 0.993151
Scruz       -0.240524   0.215402  -1.117 0.275208
Adjacent    -0.074805   0.017700  -4.226 0.000297

(Intercept)
Area
Elevation   ***
Nearest
Scruz
Adjacent    ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,    Adjusted R-squared:  0.7171
F-statistic:  15.7 on 5 and 24 DF,  p-value: 6.838e-07
```

Multiple Linear Regression in Matrix Notation

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & X_{1,1} & X_{2,1} & \cdots & X_{p-1,1} \\ 1 & X_{1,2} & X_{2,2} & \cdots & X_{p-1,2} \\ \vdots & \dots & \ddots & \vdots & \\ 1 & X_{1,n} & X_{2,n} & & X_{p-1,n} \end{pmatrix}$$

We can express MLR as

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_{p-1})^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)^T$

Error Sum of Squares (SSE) $= \sum_{i=1}^{n}(Y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j X_j)^2$ can be expressed in Matrix notation as:

$$(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$$

Multiple Linear Regression Topics

Similar to SLR, we will discuss

- Estimation
- Inference
- Diagnostics and Remedies

We will also discuss some new topics

- Model Selection
- Multicollinearity