Lecture 24

Simple Linear Regression III

Readings: IntroStat Chapter 11; OpenIntro Chapter 8

STAT 8010 Statistical Methods I
June 20, 2023

Whitney Huang
Clemson University

Agenda

1. Correlation and Simple Linear Regression

2. Advanced Topics in Regression Analysis

Correlation and Simple Linear Regression

- Pearson Correlation: $r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$
- $-1 \leq r \leq 1$ measures the strength of the **linear relationship** between Y and X
- We can show

$$r = \hat{\beta}_1 \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}},$$

this implies

$$\beta_1 = 0 \text{ in SLR} \Leftrightarrow \rho = 0$$

Coefficient of Determination R^2

- Defined as the proportion of total variation explained by SLR

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- We can show $r^2 = R^2$:

$$\begin{aligned} r^2 &= \left(\hat{\beta}_1 \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \right)^2 \\ &= \frac{\hat{\beta}_1^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= \frac{\text{SSR}}{\text{SST}} \\ &= R^2 \end{aligned}$$

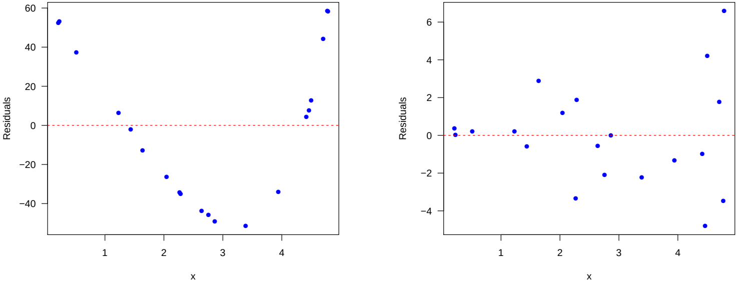
Maximum Heart Rate vs. Age: r and R^2


```
> summary(fit)$r.squared
[1] 0.9090967
> cor(Age, MaxHeartRate)
[1] -0.9534656
```

Interpretation:

There is a strong negative linear relationship between `MaxHeartRate` and `Age`. Furthermore, $\sim 91\%$ of the variation in `MaxHeartRate` can be explained by `Age`.

SLR Model Remedies



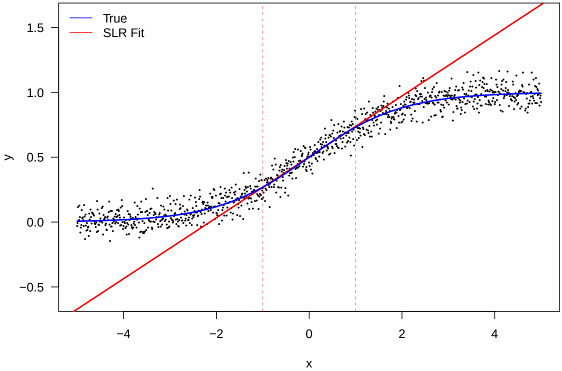
⇒ Nonlinear relationship

- Transform X
- Nonlinear regression

⇒ Non-constant variance

- Transform Y
- Weighted least squares

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

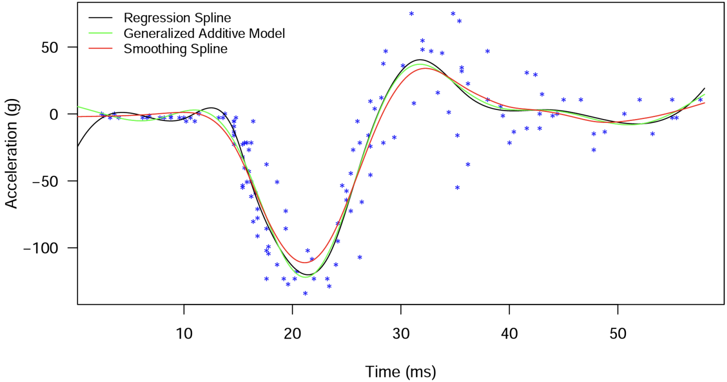
- **Model:** $Y = \beta_0 + \beta_1 X + \varepsilon, \qquad \varepsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$
- **Estimation:** Use the method of least squares to estimate the parameters $(\beta_0, \beta_1, \sigma^2)$
- **Inference**
  - **Hypothesis Testing**
  - **Confidence/prediction Intervals**
  - **ANOVA**
- **Model Diagnostics and Remedies**

Advanced Topics

Non-parametric Regression

$$Y = f(x) + \varepsilon \Rightarrow \mathrm{E}[Y|x] = f(x),$$

where $f(x)$ is a smooth function estimated from the data

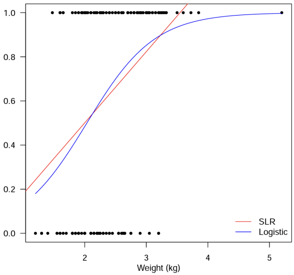


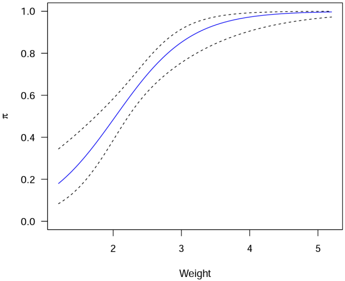
Logistic Regression

Y: binary response with the "success" probability π

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x.$$

- $\log(\frac{\pi}{1-\pi})$: the log-odds or the logit
- $\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \in (0,1)$

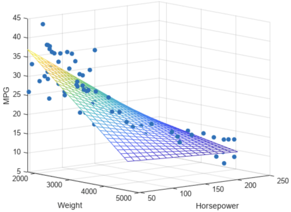




Multiple Linear Regression

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

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon_i, \quad \varepsilon_i \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma^2)$$



Source: https://www.mathworks.com/help/stats/regress.html

New Topics:

- Model Selection
- Multicollinearity

Analysis of Covariance (ANCOVA)

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_{p-1} x_{p-1} + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2)$$

$x_1, x_2, \cdots, x_{p-1}$ are the predictors.

ANCOVA is a statistical method used to handle situations where some of the predictors involve qualitative (categorical) variables