Lecture 37

Simple Linear Regression: Resiual Analysis and Hypothesis Testing

STAT 8010 Statistical Methods I
November 29, 2019

Whitney Huang
Clemson University

Agenda

1. **Review of Last Class**
2. **Residual Analysis**
3. **Hypothesis Testing**

Simple Linear Regression (SLR)

Y: dependent (response) variable; X: independent (predictor) variable

- In SLR we **assume** there is a **linear relationship** between X and Y:

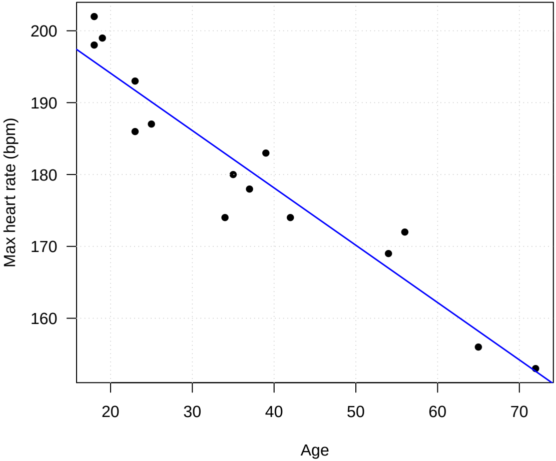
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

  where $\mathrm{E}(\varepsilon_i) = 0$, and $\mathrm{Var}(\varepsilon_i) = \sigma^2, \forall i$. Furthermore, $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$

- **Least Squares Estimation:** $\mathrm{argmin}_{\beta_0,\beta_1} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \Rightarrow$
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  - $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$

- **Residuals**: $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_{0,\text{LS}} + \hat{\beta}_{1,\text{LS}} X_i$

Maximum Heart Rate vs. Age: SLR Fit



Question: Is linear relationship between max heart rate and age reasonable? ⇒ Residual Analysis

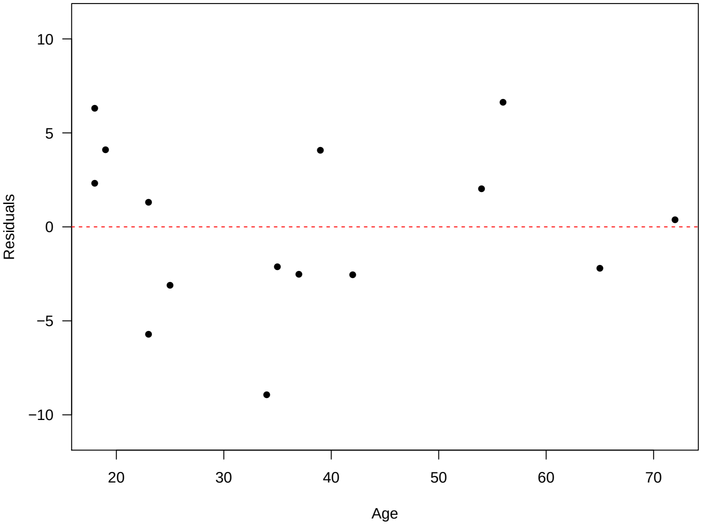
Residuals

- The residuals are the differences between the observed and fitted values:
$$e_i = Y_i - \hat{Y}_i,$$
where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

- e_i is NOT the error term $\varepsilon_i = Y_i - \mathrm{E}[Y_i]$

- Residuals are very useful in assessing the appropriateness of the assumptions on ε_i. Recall
  - $\mathrm{E}[\varepsilon_i] = 0$
  - $\mathrm{Var}[\varepsilon_i] = \sigma^2$
  - $\mathrm{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$

Maximum Heart Rate vs. Age Residual Plot: ε vs. X



Interpreting Residual Plots

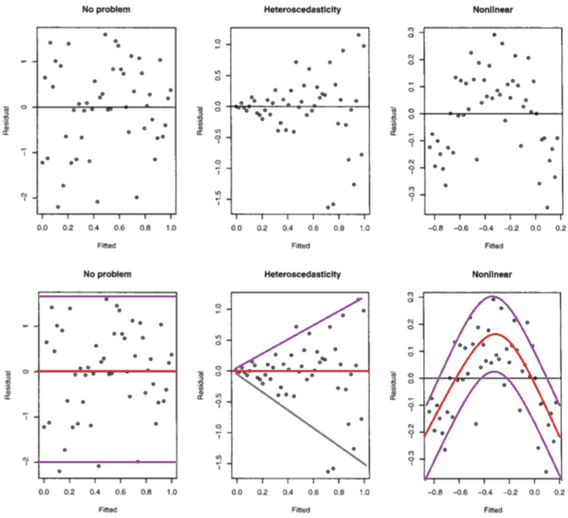
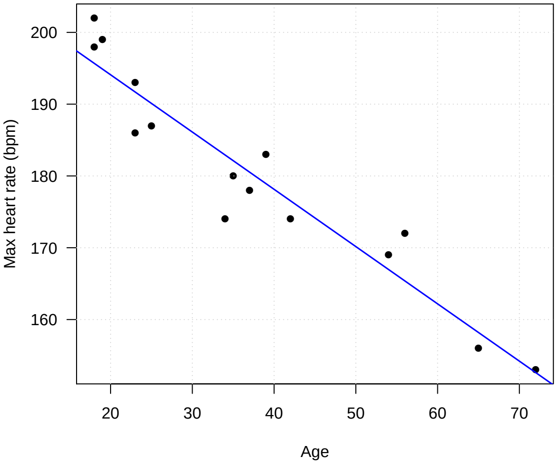


Figure: Figure courtesy of Faraway's Linear Models with R (2005, p. 59).

How (Un)certain We Are?



Can we formally quantify our estimation uncertainty? $\Rightarrow$
We need additional (distributional) assumption on ε

Normal Error Regression Model

Recall

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Further assume $\varepsilon_i \sim \mathrm{N}(0, \sigma^2) \Rightarrow Y_i \sim \mathrm{N}(\beta_0 + \beta_1 X_i, \sigma^2)$
- With normality assumption, we can derive the **sampling distribution** of $\hat{\beta}_1$ and $\hat{\beta}_0 \Rightarrow$
  - $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}, \quad \hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$
  - $\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2}, \quad \hat{\sigma}_{\hat{\beta}_0} = \hat{\sigma}\sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$

  where t_{n-2} denotes the Student's t distribution with $n-2$ degrees of freedom

Maximum Heart Rate vs. Age: Hypothesis Test for Slope

1. $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
2. Compute the **test statistic**: $t^* = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{-0.7977}{0.06996} = -11.40$
3. Compute **P-value**: $\mathrm{P}(|t_{13}| \geq |t^*|) = 3.85 \times 10^{-8}$
4. Compare to α and draw conclusion:

> Reject H_0 at α = .05 level, evidence suggests a negative linear relationship between `MaxHeartRate` and `Age`

Maximum Heart Rate vs. Age: Hypothesis Test for Intercept

1. $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq= 0$
2. Compute the **test statistic**: $t^* = \frac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\beta_0}} = \frac{210.0485}{2.86694} = 73.27$
3. Compute **P-value**: $\mathrm{P}(|t_{13}| \geq |t^*|) \simeq 0$
4. Compare to α and draw conclusion:

> Reject H_0 at $\alpha = .05$ level, evidence suggests evidence suggests the intercept (the expected `MaxHeartRate` at age 0) is different from 0

Summary

In this lecture, we learned

- Residual analysis to (graphically) check model assumptions
- **Normal Error Regression Model** and statistical inference for β_0 and β_1

Next time we will talk about

1. Confidence/Prediction Intervals
2. Analysis of Variance (ANOVA) Approach to Regression