

# Lecture 25

## Simple Linear Regression: Model Assumptions & Inferences

Text: Chapter 11

*STAT 8010 Statistical Methods I*

November 24, 2020

Whitney Huang  
Clemson University

# Linear Regression Fit

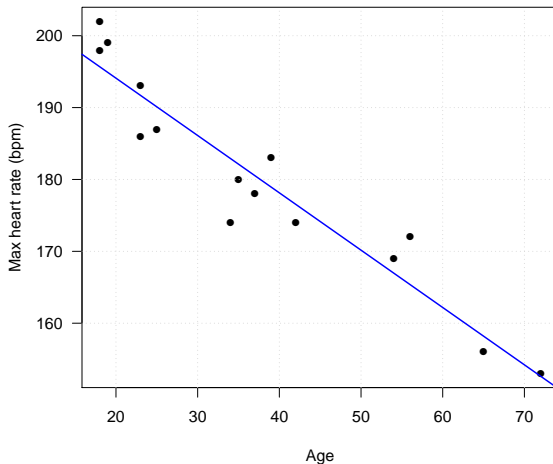
Simple Linear  
Regression: Model  
Assumptions &  
Inferences



Residual Analysis

Confidence/Prediction  
Intervals

Hypothesis Testing



**Question:** Is linear relationship between max heart rate and age reasonable?  $\Rightarrow$  [Residual Analysis](#)

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

# Maximum Heart Rate vs. Age Residual Plot: $\varepsilon$ vs. $X$

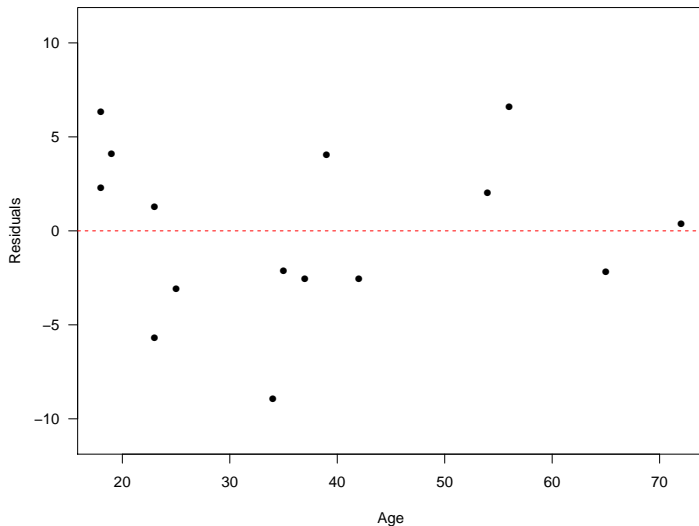
Simple Linear  
Regression: Model  
Assumptions &  
Inferences



Residual Analysis

Confidence/Prediction  
Intervals

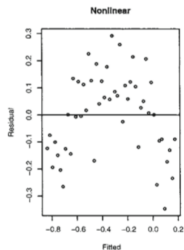
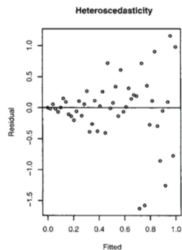
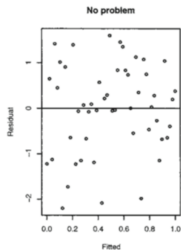
Hypothesis Testing



# Interpreting Residual Plots

Simple Linear  
Regression: Model  
Assumptions &  
Inferences

CLEMSON  
UNIVERSITY

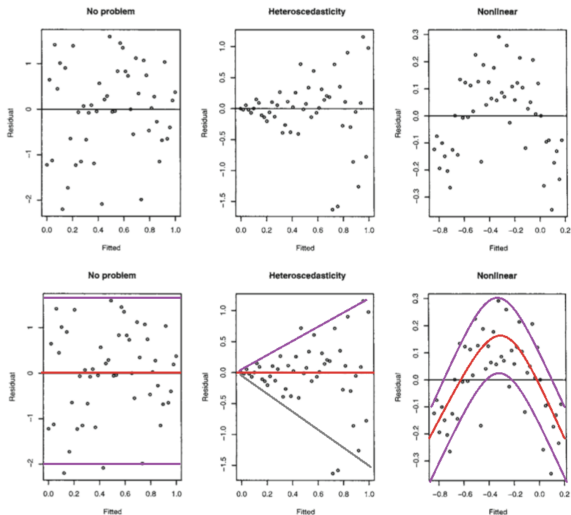


Residual Analysis

Confidence/Prediction  
Intervals

Hypothesis Testing

# Interpreting Residual Plots



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

## Recap: Simple Linear Regression

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

- **Least Squares Estimation:**

$$\text{argmin}_{\beta_0, \beta_1} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \Rightarrow$$

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

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

- $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$

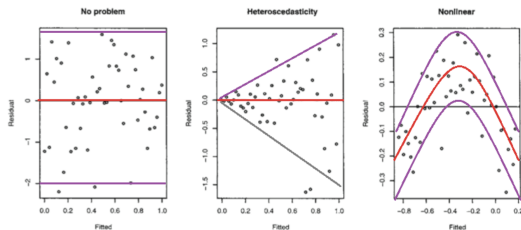
- **Residuals:**  $e_i = Y_i - \hat{Y}_i$ , where  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

## Recap: Residual Analysis

- **Residual Analysis:** To check the appropriateness of SLR model

- Is the regression function linear?
- Do  $\varepsilon_i$ 's have constant variance  $\sigma^2$ ?
- Are  $\varepsilon_i$ 's independent to each other?

We plot residuals  $e_i$ 's against  $X_i$ 's (or  $\hat{Y}_i$ 's) to assess these aspects



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



# How (Un)certain We Are?

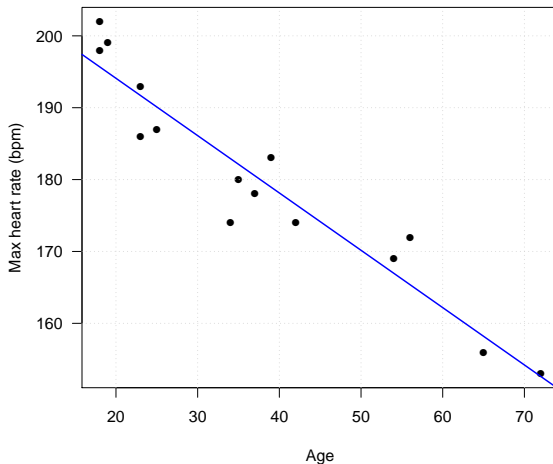
Simple Linear  
Regression: Model  
Assumptions &  
Inferences

CLEMSON  
UNIVERSITY

Residual Analysis

Confidence/Prediction  
Intervals

Hypothesis Testing



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

## Recall

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

- Further assume  $\varepsilon_i \sim N(0, \sigma^2) \Rightarrow Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$
- 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

- Recall  $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$ , we use this fact to construct **confidence intervals (CIs)** for  $\beta_1$ :

$$\left[ \hat{\beta}_1 - t_{\alpha/2, n-2} \hat{\sigma}_{\hat{\beta}_1}, \hat{\beta}_1 + t_{\alpha/2, n-2} \hat{\sigma}_{\hat{\beta}_1} \right],$$

where  $\alpha$  is the **confidence level** and  $t_{\alpha/2, n-2}$  denotes the  $1 - \alpha/2$  percentile of a student's t distribution with  $n - 2$  degrees of freedom

- Similarly, we can construct CIs for  $\beta_0$ :

$$\left[ \hat{\beta}_0 - t_{\alpha/2, n-2} \hat{\sigma}_{\hat{\beta}_0}, \hat{\beta}_0 + t_{\alpha/2, n-2} \hat{\sigma}_{\hat{\beta}_0} \right]$$

## Interval Estimation of $E(Y_h)$

- We often interested in estimating the **mean** response for a particular value of predictor, say,  $X_h$ . Therefore we would like to construct CI for  $E[Y_h]$

- We need sampling distribution of  $\hat{Y}_h$  to form CI:

- $\frac{\hat{Y}_h - Y_h}{\hat{\sigma}_{\hat{Y}_h}} \sim t_{n-2}, \quad \hat{\sigma}_{\hat{Y}_h} = \hat{\sigma} \sqrt{\left( \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)}$

- CI:

$$\left[ \hat{Y}_h - t_{\alpha/2, n-2} \hat{\sigma}_{\hat{Y}_h}, \hat{Y}_h + t_{\alpha/2, n-2} \hat{\sigma}_{\hat{Y}_h} \right]$$

- **Quiz:** Use this formula to construct CI for  $\beta_0$

- Suppose we want to predict the response of a future observation given  $X = X_h$
- We need to account for added variability as a new observation does not fall directly on the regression line (i.e.,  $Y_{h(\text{new})} = E[Y_h] + \varepsilon_h$ )
- Replace  $\hat{\sigma}_{\hat{Y}_h}$  by  $\hat{\sigma}_{\hat{Y}_{h(\text{new})}} = \hat{\sigma} \sqrt{\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)}$  to construct CIs for  $Y_{h(\text{new})}$

## Maximum Heart Rate vs. Age Revisited

The maximum heart rate  $\text{MaxHeartRate}$  ( $\text{HR}_{\max}$ ) of a person is often said to be related to age  $\text{Age}$  by the equation:

$$\text{HR}_{\max} = 220 - \text{Age}.$$

Suppose we have 15 people of varying ages are tested for their maximum heart rate (bpm)

Age	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
$\text{HR}_{\max}$	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

- Construct the 95% CI for  $\beta_1$
- Compute the estimate for mean  $\text{MaxHeartRate}$  given  $\text{Age} = 40$  and construct the associated 90% CI
- Construct the prediction interval for a new observation given  $\text{Age} = 40$

# 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**:  $P(|t^*| \geq |t_{obs}|) = 3.85 \times 10^{-8}$
- 4 Compare to  $\alpha$  and draw conclusion:

Reject  $H_0$  at  $\alpha = .05$  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**:  $P(|t^*| \geq |t_{obs}|) \simeq 0$

4 Compare to  $\alpha$  and draw conclusion:

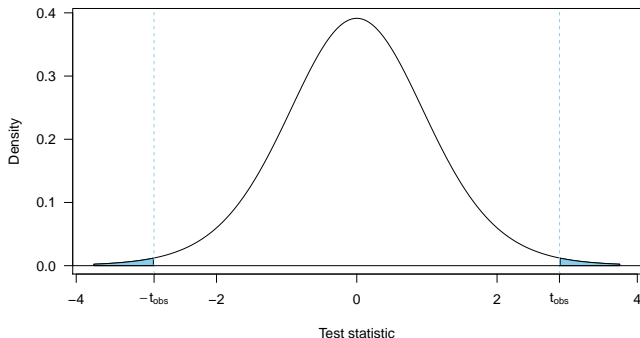
Reject  $H_0$  at  $\alpha = .05$  level, evidence suggests evidence suggests the intercept (the expected `MaxHeartRate` at age 0) is different from 0



## Hypothesis Tests for $\beta_{\text{age}} = -1$

$$H_0 : \beta_{\text{age}} = -1 \text{ vs. } H_a : \beta_{\text{age}} \neq -1$$

$$\text{Test Statistic: } \frac{\hat{\beta}_{\text{age}} - (-1)}{\hat{\sigma}_{\hat{\beta}_{\text{age}}}} = \frac{-0.79773 - (-1)}{0.06996} = 2.8912$$



Residual Analysis

Confidence/Prediction  
Intervals

Hypothesis Testing

$$\text{P-value: } 2 \times \mathbb{P}(t^* > 2.8912) = 0.013, \text{ where } t^* \sim t_{df=13}$$

In this lecture, we learned

- **Residual Analysis** for checking model assumptions
- **Normal Error Regression Model** and **statistical inference** for  $\beta_0$  and  $\beta_1$
- **Confidence/Prediction Intervals & Hypothesis Testing**

Next time we will talk about

- 1 Analysis of Variance (ANOVA) Approach to Regression
- 2 Correlation ( $r$ ) & Coefficient of Determination ( $R^2$ )