Lecture 3

Simple Linear Regression II

Reading: Chapter 11

STAT 8020 Statistical Methods II August 26, 2019

> Whitney Huang Clemson University



Agenda

- Review of Last Class
- Residual Analysis
- 3 Hypothesis Testing



Notes

Notes

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, $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$

Least Squares Estimation:

 $\begin{array}{l} \operatorname{argmin}_{\beta_0,\beta_1} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \Rightarrow \\ \bullet \ \hat{\beta}_{1,\mathrm{LS}} = \frac{\sum_{i=1}^n (X_i - X_i)(Y_i - Y)}{\sum_{i=1}^n (X_i - X)^2} \end{array}$

- $\bullet \ \hat{\beta}_{0,\mathsf{LS}} = \bar{\mathit{Y}} \hat{\beta}_{1,\mathsf{LS}}\bar{\mathit{X}}$
- $\hat{\sigma}_{LS}^2 = \frac{\sum_{i=1}^{n} (Y_i \hat{Y}_i)^2}{n-2}$
- Residuals: $e_i = Y_i \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_{0,LS} + \hat{\beta}_{1,LS} X_i$

Regression II					
CLEMS#N					
Review of Last Class					

Simple Linear

Notes

Maximum Heart Rate vs. Age

The maximum heart rate MaxHeartRate of a person is often said to be related to age Age by the equation:

$$\label{eq:maxHeartRate} {\sf MaxHeartRate} = 220 - {\sf Age}.$$

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

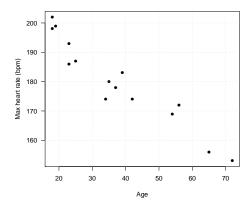
```
        Age
        18
        23
        25
        35
        65
        54
        34
        56
        72
        19
        23
        42
        18
        39
        37

        MaxHeartRate
        202
        186
        187
        180
        156
        169
        174
        172
        153
        199
        193
        174
        198
        183
        178
```

Link to this dataset: http://whitneyhuang83.
github.io/maxHeartRate.csv



Plot the Data





Notes

Notes

Estimate the Parameters β_1 , β_0 , and σ^2

 $\mathit{Y_i}$ and $\mathit{X_i}$ are the Maximum Heart Rate and Age of the i^{th} individual

- To obtain $\hat{\beta}_{1,LS}$
 - Ocompute $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
 - ② Compute $Y_i \bar{Y}, X_i \bar{X},$ and $(X_i \bar{X})^2$ for each observation
 - **3** Compute $\sum_i^n (X_i \bar{X})(Y_i \bar{X})$ divived by $\sum_i^n (X_i \bar{X})^2$
- $\hat{\beta}_{0,LS}$: Compute $\bar{Y} \hat{\beta}_{1,LS}\bar{X}$
- $\hat{\sigma}_{LS}^2$
 - Ompute the fitted values: $\hat{Y}_i = \hat{\beta}_{0,LS} + \hat{\beta}_{1,LS} X_i, \quad i = 1, \cdots, n$
 - ② Compute the **residuals** $e_i = Y_i \hat{Y}_i, \quad i = 1, \dots, n$
 - Ocompute the **residual sum of squares (RSS)** = $\sum_{i=1}^{n} (Y_i \hat{Y}_i)^2$ and divided by n 2 (why?)

Regression II
CLEMS # N
UNIVERSITY
Review of Last Class

Notes			

Let's Do the Calculations

$$\bar{X} = \sum_{i=1}^{15} \frac{18 + 23 + \dots + 39 + 37}{15} = 37.33$$

$$\bar{Y} = \sum_{i=1}^{15} \frac{202 + 186 + \dots + 183 + 178}{15} = 180.27$$

$$\hat{\beta}_{1,\text{LS}} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = 0.7977$$

$$\hat{eta}_{0, extsf{LS}} = \bar{Y} - \hat{eta}_{1, extsf{LS}} \bar{X} = 210.0485$$

$$\hat{\sigma}_{\text{LS}}^2 = \frac{\sum_{i=1}^{15} (Y_i - \hat{Y}_i)^2}{13} = 20.9563 \Rightarrow \hat{\sigma} = 4.5778$$

Notes

Let's Double Check

> fit <- lm(MaxHeartRate ~ Age)
> summary(fit)
내게 입니다 그러지 않는데 가지 않는데 그렇게 하는데 하는데 그렇게 되었다.
Call:
<pre>lm(formula = MaxHeartRate ~ Age)</pre>
[18] 이번 사람이 사용하는 이번 사용하는 사용하는 사용하는 사용하는 사용하는 사용하는 사용하는 사용하는
Residuals:
Min 1Q Median 3Q Max
-8.9258 -2.5383 0.3879 3.1867 6.6242
CCC:-ib
Coefficients:
Estimate Std. Error t value Pr(> t)
(Intercept) 210.04846
Age -0.79773 0.06996 -11.40 3.85e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 4.578 on 13 degrees of freedom
Multiple R-squared: 0.9091, Adjusted R-squared: 0.9021



Review of Last Class Residual Analysis Hypothesis Testing

3.8

Notes

Output from imp

Load the data

Parameter Estimates

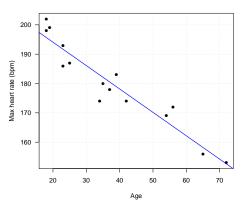
Term Estimate Std Error t Ratio Prob>|t|
Intercept 210.04846 2.866939 73.27 <.0001*
Age -0.797727 0.069963 -11.40 <.0001*

Simple Linear Regression II								
	C	Ē	N	43	St	í	1	1
I N		٧	ε	R	S		T	Y

Review of Last Class Residual Analysis

Notes			
-			

Linear Regression 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:

$$e_i = Y_i - \hat{Y}_i,$$

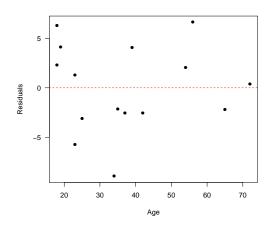
where
$$\hat{Y}_i = \hat{\beta}_{0,LS} + \hat{\beta}_{1,LS} X_i$$

- ullet e_i is NOT the error term $arepsilon_i = Y_i \mathrm{E}[Y_i]$
- Residuals are very useful in assessing the appropriateness of the assumptions on ε_i . Recall
 - $\bullet \ E[\varepsilon_i] = 0$
 - $Var[\varepsilon_i] = \sigma^2$
 - $\operatorname{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$



Notes

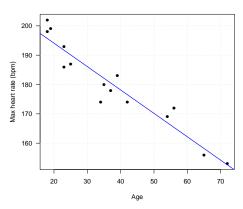
Residual Plot: ε vs. X



Simple Linear Regression II							
CLEMS N							
Residual Analysis							

Notes

How (Un)certain We Are?



Can we formally quantify our estimation uncertainty?

 \Rightarrow We need additional (distributional) assumption on ε



Notes

Normal Error Regression Model

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 $\hat{\beta}_0 \Rightarrow$

$$\begin{array}{ll} \bullet & \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}} \sim t_{n-2}, \quad \hat{\sigma}_{\beta_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2}} \\ \bullet & \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\beta_0}} \sim t_{n-2}, \quad \hat{\sigma}_{\beta_0} = \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)} \end{array}$$

 $\begin{array}{ll} & \stackrel{\cdot}{\beta_0-\beta_0} \sim t_{n-2}, & \hat{\sigma}_{\beta_0} = \hat{\sigma}\sqrt{(\frac{1}{n} + \frac{\tilde{\chi}^2}{\sum_{i=1}^n(X_i-X)^2})}\\ \text{where } t_{n-2} \text{ denotes the Student's t distribution with } \\ n-2 \text{ degrees of freedom} \end{array}$



Notes

1 $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 \neq = 0$

Steps of Hypothesis Test for Slope

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

Reject H_0 at $\alpha = .05$ level, evidence suggests a negative linear relationship between MaxHeartRate and Age



Notes			
-			

Steps of Hypothesis Test for Intercept

1 $H_0: \beta_0 = 0$ vs. $H_a: \beta_0 \neq = 0$

② Compute the **test statistic**: $t^* = \frac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\beta_0}} = \frac{210.0485}{2.86694} = 73.27$

Our Compute P-value: $P(|t_{13}| \ge |t^*|) \simeq 0$

lacktriangle Compare to α and draw conclusion:

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

Simple Linear Regression II
CLEMS N
Hypothesis Testing
3.16

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

- Confidence/Prediction Intervals
- Analysis of Variance (ANOVA) Approach to Regression



Notes _____

Notes ______