Lecture 3

Simple Linear Regression II

Reading: Chapter 11

STAT 8020 Statistical Methods II
August 26, 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:** $\mathrm{argmin}_{\beta_0,\beta_1} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \Rightarrow$
  - $\hat{\beta}_{1,\mathrm{LS}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
  - $\hat{\beta}_{0,\mathrm{LS}} = \bar{Y} - \hat{\beta}_{1,\mathrm{LS}} \bar{X}$
  - $\hat{\sigma}^2_{\mathrm{LS}} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$
- **Residuals**: $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_{0,\mathrm{LS}} + \hat{\beta}_{1,\mathrm{LS}} X_i$

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:

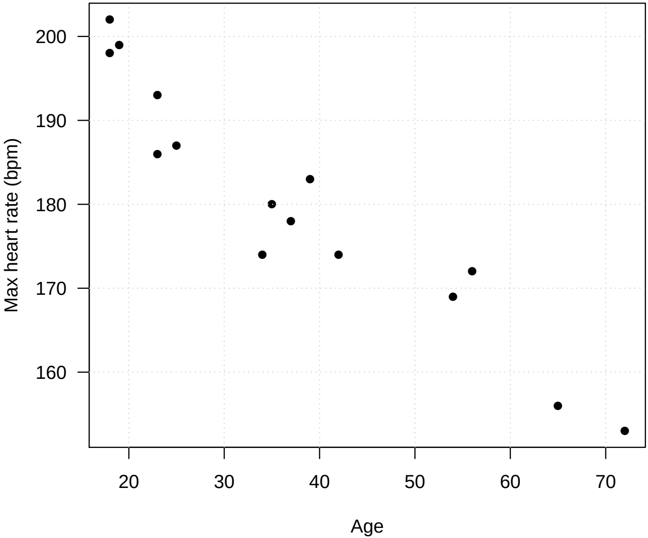
$$\text{MaxHeartRate} = 220 - \text{Age}.$$

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
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



Estimate the Parameters β_1, β_0, and σ^2

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

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

Let's Do the Calculations

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

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

X	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
Y	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178
$X-\bar{X}$	-19.33	-14.33	-12.33	-2.33	27.67	16.67	-3.33	18.67	34.67	-18.33	-14.33	4.67	-19.33	1.67	-0.33
$Y-\bar{Y}$	21.73	5.73	6.73	-0.27	-24.27	-11.27	-6.27	-8.27	-27.27	18.73	12.73	-6.27	17.73	2.73	-2.27
$(X-\bar{X})(Y-\bar{Y})$	-420.18	-82.18	-83.04	0.62	-671.38	-187.78	20.89	-154.31	-945.24	-343.44	-182.51	-29.24	-342.84	4.56	0.76
$(X-\bar{X})^2$	373.78	205.44	152.11	5.44	765.44	277.78	11.11	348.44	1201.78	336.11	205.44	21.78	373.78	2.78	0.11
$\hat{Y}$	195.69	191.70	190.11	182.13	158.20	166.97	182.93	165.38	152.61	194.89	191.70	176.54	195.69	178.94	180.53

- $\hat{\beta}_{1,\text{LS}} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = 0.7977$
- $\hat{\beta}_{0,\text{LS}} = \bar{Y} - \hat{\beta}_{1,\text{LS}}\bar{X} = 210.0485$
- $\hat{\sigma}^2_{\text{LS}} = \frac{\sum_{i=1}^{15}(Y_i-\hat{Y}_i)^2}{13} = 20.9563 \Rightarrow \hat{\sigma} = 4.5778$

Let's Double Check

Output from R (RStudio)

```
> fit <- lm(MaxHeartRate ~ Age)
> summary(fit)

Call:
lm(formula = MaxHeartRate ~ Age)

Residuals:
    Min      1Q  Median      3Q     Max
-8.9258 -2.5383  0.3879  3.1867  6.6242

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 210.04846    2.86694   73.27  < 2e-16 ***
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
F-statistic:   130 on 1 and 13 DF,  p-value: 3.848e-08
```

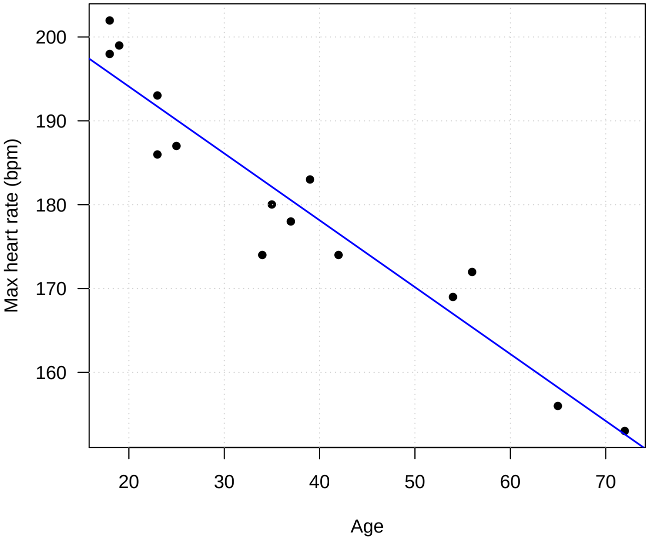
Output from jmp

1. Load the data
2. `Analyze` → `Fit Model` → `Run`

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*

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:

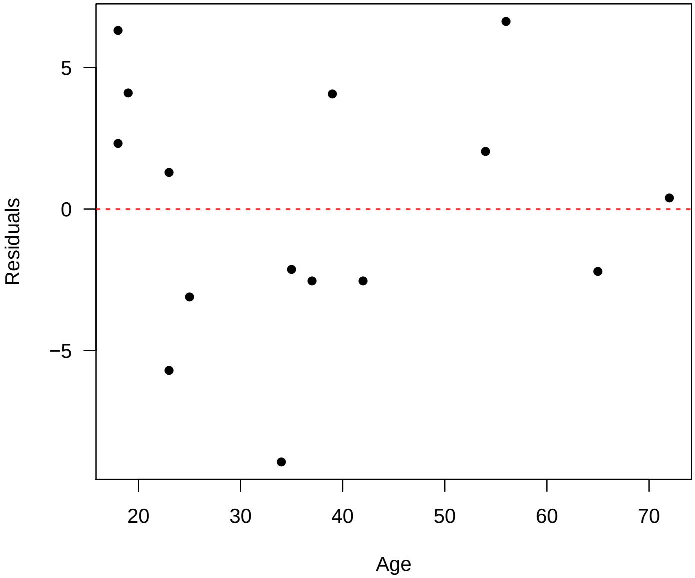
$$e_i = Y_i - \hat{Y}_i,$$

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

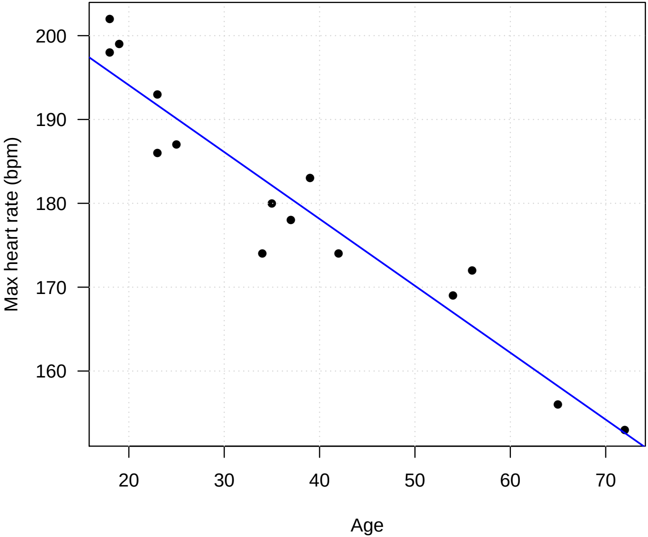
- e_i is NOT the error term $\varepsilon_i = Y_i - \mathrm{E}[Y_i]$

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

Residual Plot: ε vs. X



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

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

Steps of Hypothesis Test for Slope

1. $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq= 0$
2. Compute the **test statistic**: $t^* = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\beta_1}} = \frac{-0.7977}{0.06996} = -11.40$
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`

Steps of Hypothesis Test for Intercept

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

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

3. Compute **P-value**: $\mathrm{P}(|t_{13}| \geq |t^*|) \simeq 0$

4. 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

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