

Lecture 1

Review of Simple Linear Regression

DSA 8020 Statistical Methods II
January 6, 2021

Simple Linear
Regression

Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

Whitney Huang
Clemson University

- 1 Simple Linear Regression
- 2 Parameter Estimation
- 3 Residual Analysis
- 4 Confidence/Prediction Intervals
- 5 Hypothesis Testing

Simple Linear
Regression

Parameter Estimation

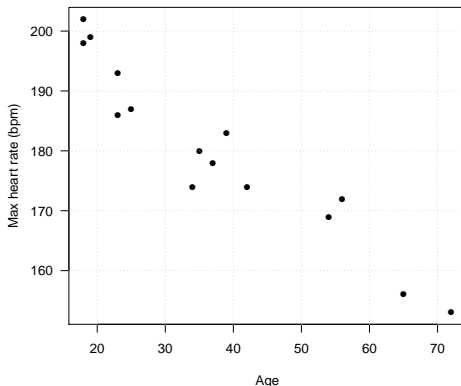
Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

What is Regression Analysis?

Regression analysis: A set of statistical procedures for estimating the relationship between **response variable** and **predictor variable(s)**



Simple linear regression: The relationship between the response variable and the predictor variable is approximately linear

Y : response variable; x : predictor variable

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

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

- We need to estimate β_0 (intercept) and β_1 (slope) based on observed data $\{x_i, y_i\}_{i=1}^n$
- We can use the estimated regression equation to
 - make predictions
 - study the relationship between response and predictor
 - control the response
- Yet we need to quantify our **estimation uncertainty** regarding the linear relationship

Simple Linear
Regression

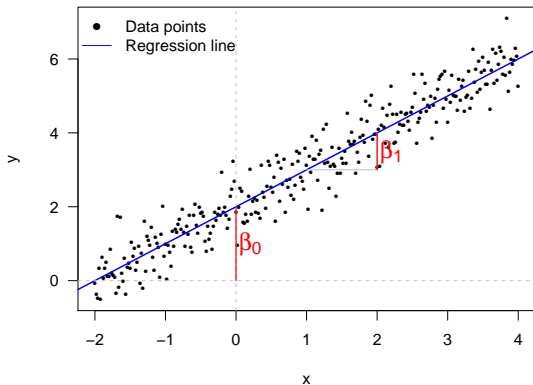
Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

Regression equation: $Y = \beta_0 + \beta_1 x$



● β_0 : $E[Y]$ when $x = 0$

● β_1 : $E[\Delta Y]$ when x increases by 1

Assumptions about the Random Error ε

In order to estimate β_0 and β_1 , we make the following assumptions about ε

- $E[\varepsilon_i] = 0$
- $\text{Var}[\varepsilon_i] = \sigma^2$
- $\text{Cov}[\varepsilon_i, \varepsilon_j] = 0, \quad i \neq j$

Therefore, we have

$$E[Y_i] = \beta_0 + \beta_1 x_i, \text{ and}$$

$$\text{Var}[Y_i] = \sigma^2$$

The regression line $\beta_0 + \beta_1 x$ represents the **conditional mean curve** whereas σ^2 measures the magnitude of the **variation** around the regression curve

Estimation: Method of Least Square

For given observations $(x_i, y_i)_{i=1}^n$, choose β_0 and β_1 to minimize the *sum of squared errors*:

$$L(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

Solving the above minimization problem requires some knowledge from Calculus....

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

We also need to **estimate** σ^2

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}, \text{ where } \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Example: 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) (link to the “dataset”:

[whitneyhuang83.github.io/STAT8010/Data/
maxHeartRate.csv](https://github.com/whitneyhuang83/STAT8010/Data/maxHeartRate.csv))

- 1 Compute the estimates for the regression coefficients
- 2 Compute the fitted values
- 3 Compute the estimate for σ

Simple Linear
Regression

Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

Maximum Heart Rate vs. Age

Output from  ( R Studio)

```
> 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

Simple Linear
Regression

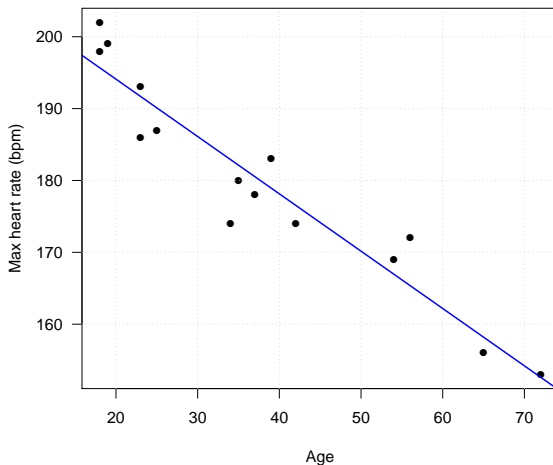
Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

Assessing Linear Regression Fit



Simple Linear
Regression

Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

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

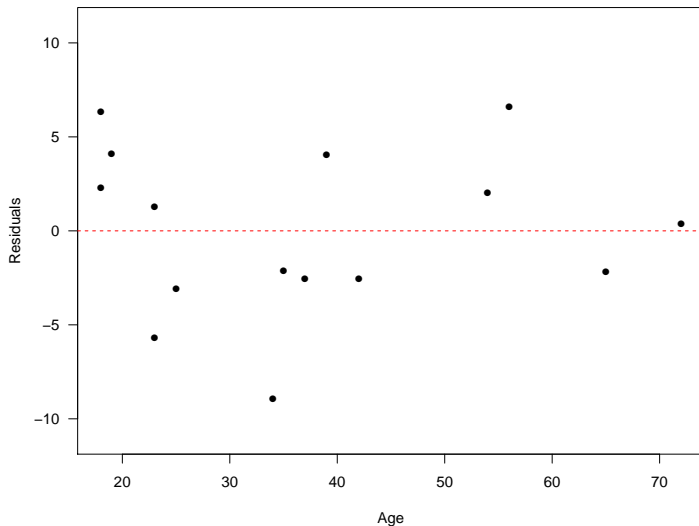
- The **residuals** are the differences between the observed and fitted values:

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

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

- Residuals are very useful in assessing the appropriateness of the assumptions on ε_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: ε vs. x



Simple Linear
Regression

Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

Interpreting Residual Plots

Simple Linear
Regression

Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing

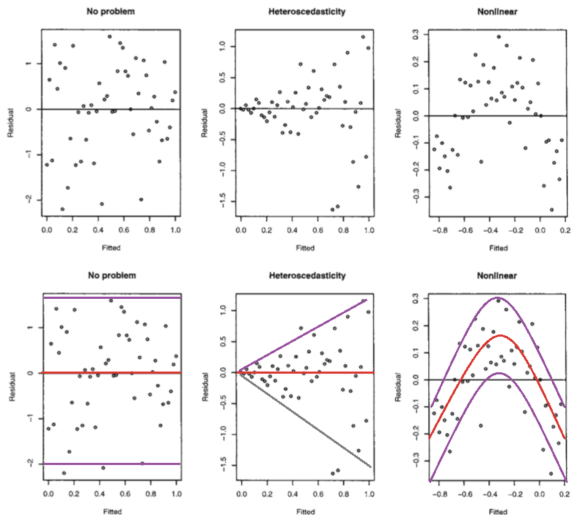


Figure: Figure courtesy of Faraway's Linear Models with R (2005, p. 59).

How (Un)certain We Are?

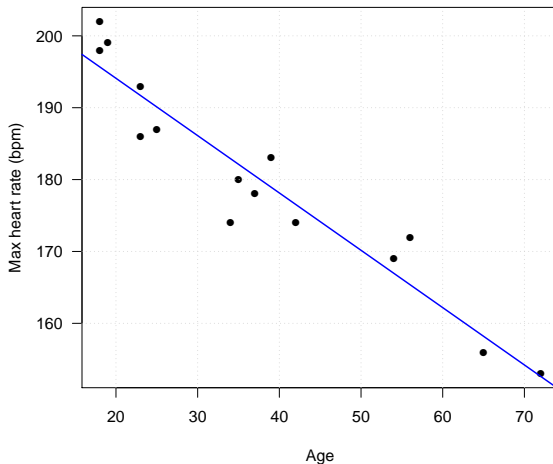
Simple Linear
Regression

Parameter Estimation

Residual Analysis

Confidence/Prediction
Intervals

Hypothesis Testing



Can we formally quantify our estimation uncertainty? \Rightarrow

We need additional (distributional) assumption on ε

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 $\hat{\beta}_0 \Rightarrow$

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

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

Simple Linear
Regression

Parameter Estimation

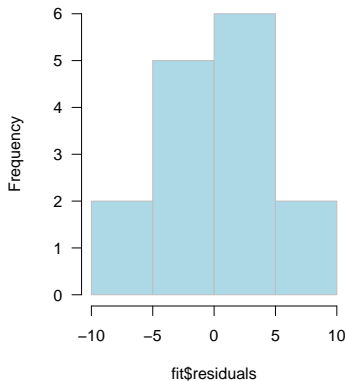
Residual Analysis

Confidence/Prediction
Intervals

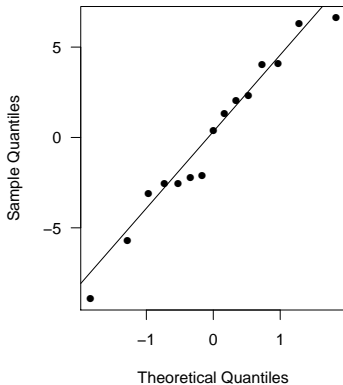
Hypothesis Testing

Assessing Normality Assumption on ε

Histogram of fit\$residuals



Normal Q-Q Plot



- Recall $\frac{\hat{\beta}_1 - \beta_1}{\hat{SE}_{\hat{\beta}_1}} \sim t_{n-2}$, we use this fact to construct a **confidence interval (CI)** for β_1 :

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

where α 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 a CI for β_0 :

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

Confidence Interval of $E(Y_{new})$

- We often interested in estimating the **mean** response for an unobserved predictor value, say, x_{new} . Therefore we would like to construct CI for $E[Y_{new}]$, the corresponding **mean response**
- We need sampling distribution of \hat{Y}_{new} to form CI:

- $\frac{\hat{Y}_{new} - Y_{new}}{SE_{\hat{Y}_{new}}} \sim t_{n-2}, \quad SE_{\hat{Y}_{new}} = \hat{\sigma} \sqrt{\left(\frac{1}{n} + \frac{(x_{new} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}$

- CI:

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

- **Quiz:** Use this formula to construct CI for β_0

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

Maximum Heart Rate vs. Age Revisited

The maximum heart rate MaxHeartRate (HR_{\max}) of a person is often said to be related to age 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
HR_{\max}	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

- Construct the 95% CI for β_1
- Compute the estimate for mean 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**: $t^* = \frac{\hat{\beta}_1 - 0}{\hat{SE}_{\hat{\beta}_1}} = \frac{-0.7977}{0.06996} = -11.40$
- 3 Compute **P-value**: $P(|t^*| \geq |t_{obs}|) = 3.85 \times 10^{-8}$
- 4 Compare to α 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**: $t^* = \frac{\hat{\beta}_0 - 0}{\hat{SE}_{\beta_0}} = \frac{210.0485}{2.86694} = 73.27$

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

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

In this lecture, we reviewed

- Simple Linear Regression: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- Method of Least Square for parameter estimation
- Residual analysis to check model assumptions
- Confidence/Prediction Intervals and Hypothesis Testing