Lecture 2

Simple Linear Regression II

Reading: Chapter 11

STAT 8020 Statistical Methods II

August 25, 2020

Whitney Huang
Clemson University

Agenda

1. **Parameter Estimation**
2. **Residual Analysis**
3. **Confidence/Prediction Intervals**
4. **Hypothesis Testing**

Estimation: Method of Least Square

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

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

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

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

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

Estimation: Method of Least Square

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

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

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

Estimation: Method of Least Square

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

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

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

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

Properties of Least Squares Estimates

- **Gauss-Markov** theorem states that in a linear regression these least squares estimators

  1. **Are unbiased**, i.e.,
     - $E[\hat{\beta}_1] = \beta_1; E[\hat{\beta}_0] = \beta_0$
     - $E[\hat{\sigma}^2] = \sigma^2$
  2. Have **minimum variance** among all unbiased linear estimators

Note that we do not make any distributional assumption on ε_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`)

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

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^{th} individual

- To obtain $\hat{\beta}_1$
  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 $\sum_i^n (X_i - \bar{X})(Y_i - \bar{Y})$ divived by $\sum_i^n (X_i - \bar{X})^2$
- $\hat{\beta}_0$: Compute $\bar{Y} - \hat{\beta}_1 \bar{X}$
- $\hat{\sigma}^2$
  1. Compute the fitted values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1, \cdots, n$
  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
	-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
	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
	-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
	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
	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

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = -0.7977$

- $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 210.0485$

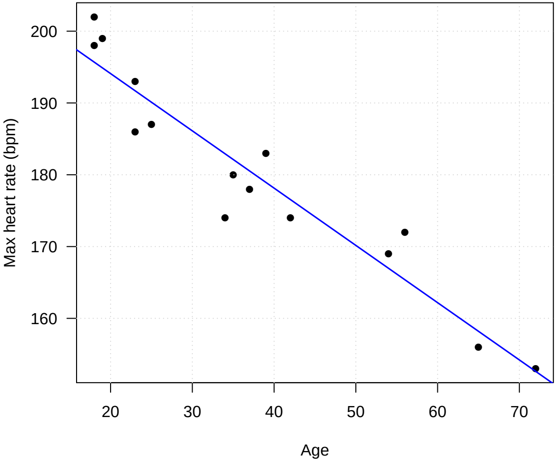
- $\hat{\sigma}^2 = \frac{\sum_{i=1}^{15}(Y_i - \hat{Y}_i)^2}{13} = 20.9563 \Rightarrow \hat{\sigma} = 4.5778$

Let's Double Check

Output from 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
```

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:

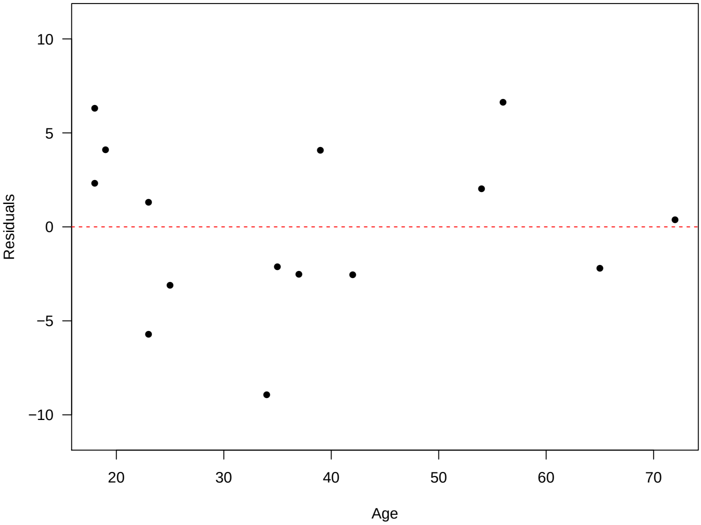
$$e_i = Y_i - \hat{Y}_i,$$

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

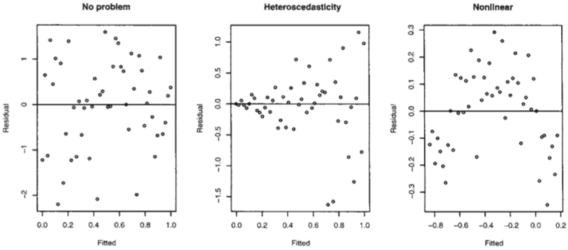
- e_i is NOT the error term $\varepsilon_i = Y_i - \mathrm{E}[Y_i]$

- 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



Interpreting Residual Plots

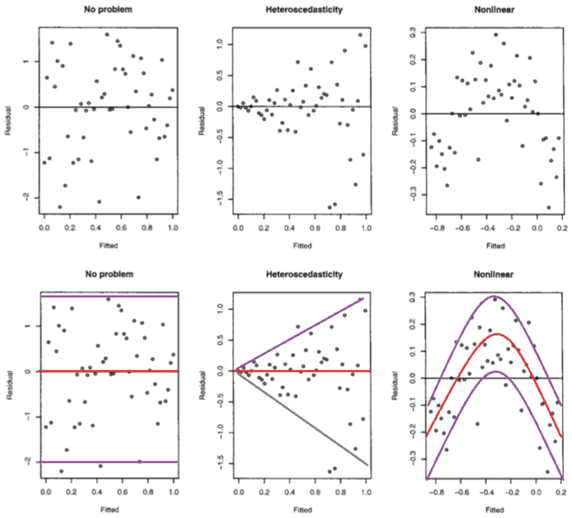
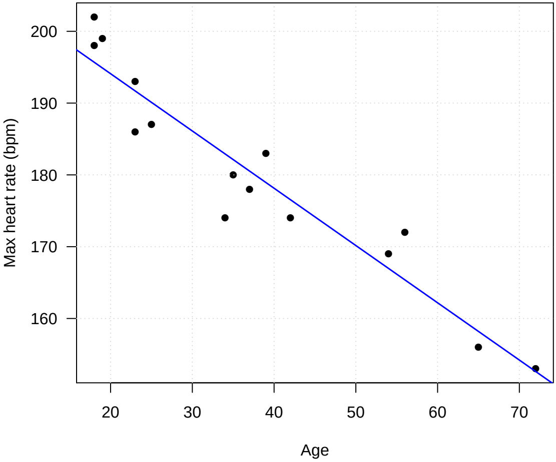


Figure: Figure courtesy of Faraway's Linear Models with R (2005, p. 59).

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

Confidence Intervals

- 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 β_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 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 CIs for β_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 $\mathrm{E}(Y_h)$

- 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 $\mathrm{E}[Y_h]$

- 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 β_0

Prediction Intervals

- 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., $Y_{\text{h(new)}} = \mathrm{E}[Y_h] + \varepsilon_h$)
- Replace $\hat{\sigma}_{\hat{Y}_h}$ by $\hat{\sigma}_{\hat{Y}_{\text{h(new)}}} = \hat{\sigma}\sqrt{\left(1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}$ to construct CIs for $Y_{\text{h(new)}}$

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:

$$HR_{max} = 220 - \text{Age}.$$

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 `Age` $= 40$ and construct the associated 90% CI
- Construct the prediction interval for a new observation given `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**: $\mathrm{P}(|t^*| \geq |t_{obs}|) = 3.85 \times 10^{-8}$
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

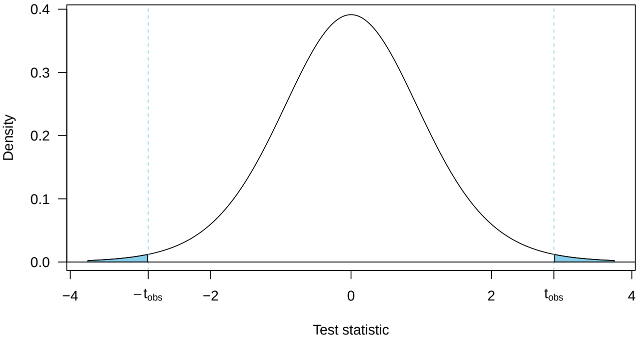
1. $H_0 : \beta_0 = 0$ vs. $H_a : \beta_0 \neq 0$
2. Compute the **test statistic**: $t^* = \frac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\beta_0}} = \frac{210.0485}{2.86694} = 73.27$
3. Compute **P-value**: $\mathrm{P}(|t^*| \geq |t_{obs}|) \simeq 0$
4. Compare to α and draw conclusion:

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

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

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

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



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

Summary

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
- statistical inference for β_0 and β_1
- Confidence/Prediction Intervals and Hypothesis Testing

Next time we will talk about

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