Lecture 5

Simple Linear Regression IV

Reading: Chapter 11

STAT 8020 Statistical Methods II

August 30, 2019

Whitney Huang
Clemson University

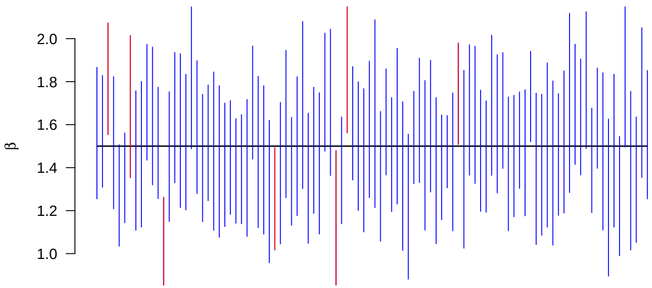
Agenda

1. **Review of Last Class**

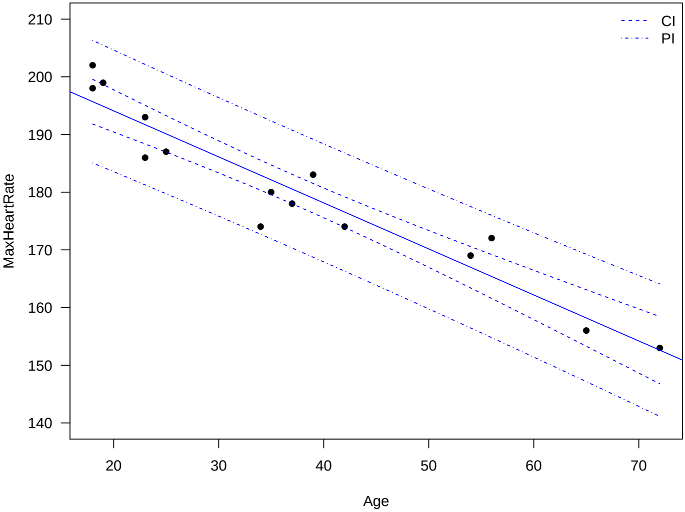
2. **Analysis of Variance (ANOVA) Approach to Regression**

Understanding Confidence Intervals

- Suppose $Y = \beta_0 + \beta_1 X + \varepsilon$, where $\beta_0 = 3$, $\beta_1 = 1.5$ and $\sigma^2 \sim \mathrm{N}(0, 1)$
- We take 100 random sample each with sample size 20
- We then construct the 95% CI for each random sample ($\Rightarrow$ 100 CIs)



Confidence Intervals vs. Prediction Intervals



Analysis of Variance (ANOVA) Approach to Regression

Partitioning Sums of Squares

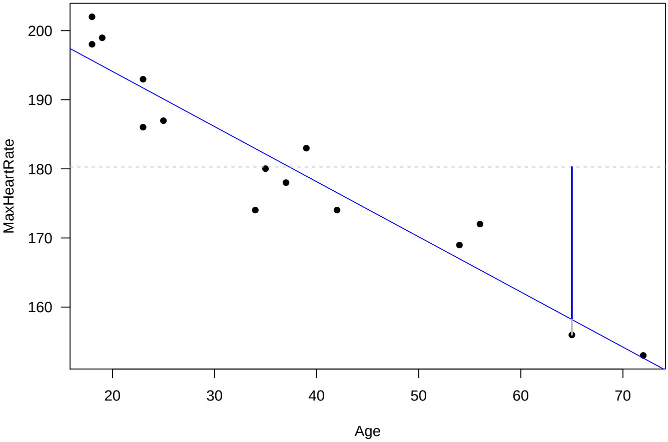
- Total sums of squares in response

$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- We can rewrite SST as

$$\begin{aligned}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 &= \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \bar{Y})^2 \\ &= \underbrace{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}_{\text{Error}} + \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{Model}}\end{aligned}$$

Partitioning Total Sums of Squares



Total Sum of Squares: SST

- If we ignored the predictor X, the $\bar{Y}$ would be the best (linear unbiased) predictor

$$Y_i = \beta_0 + \varepsilon_i \tag{1}$$

- SST is the sum of squared deviations for this predictor (i.e., $\bar{Y}$)
- The **total mean square** is SST$/(n-1)$ and represents an unbiased estimate of σ^2 under the model (1).

Regression Sum of Squares: SSR

- SSR: $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$
- Degrees of freedom is 1 due to the inclusion of the slope, i.e.,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \tag{2}$$

- “Large" MSR $=$ SSR$/1$ suggests a linear trend, because

$$\mathrm{E}[MSE] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Error Sum of Squares: SSE

- SSE is simply the sum of squared residuals

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

- Degrees of freedom is $n-2$ (Why?)

- SSE large when $|$residuals$|$ are “large" $\Rightarrow$ Y_i’s vary substantially around fitted regression line

- MSE $=$ SSE$/(n-2)$ and represents an unbiased estimate of σ^2 **when taking X into account**

ANOVA Table and F test

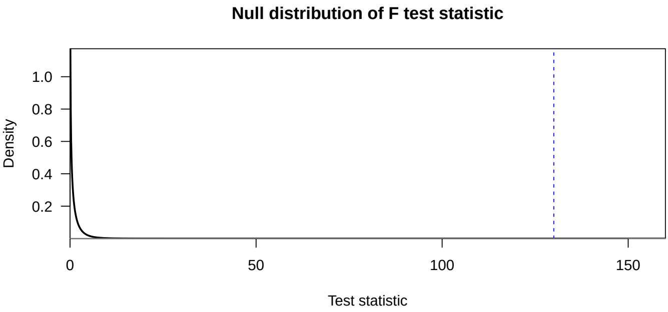
Source	df	SS	MS
Model	1	SSR $= \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$	MSR = SSR/1
Error	$n-2$	SSE $= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	MSE = SSE/(n-2)
Total	$n-1$	SST $= \sum_{i=1}^{n}(Y_i - \bar{Y})^2$	

- **Goal:** To test $H_0 : \beta_1 = 0$
- Test statistics $F^* = \frac{\text{MSR}}{\text{MSE}}$
- If $\beta_1 = 0$ then F^* should be near one $\Rightarrow$ reject H_0 when F^* “large"
- We need sampling distribution of F^* under $H_0 \Rightarrow F_{1,n-2}$, where $F(d_1, d_2)$ denotes a F distribution with degrees of freedom d_1 and d_2

F Test: $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq= 0$

```
> fit <- lm(MaxHeartRate ~ Age)
> anova(fit)
Analysis of Variance Table

Response: MaxHeartRate
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Age        1 2724.50 2724.50  130.01 3.848e-08 ***
Residuals 13  272.43   20.96                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Correlation and Simple Linear Regression

- Pearson Correlation: $r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$
- $-1 \leq r \leq 1$ measures the strength of the **linear relationship** between Y and X
- We can show $r = \hat{\beta}_{1,\text{LS}} \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$, this implies

$$\beta_1 = 0 \text{ in SLR } \Leftrightarrow \rho = 0$$

Coefficient of Determination R^2

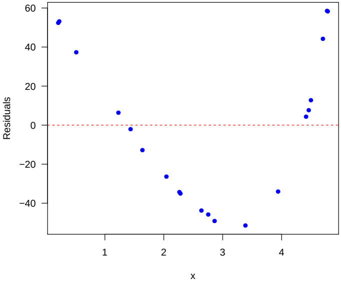
- Defined as the proportion of total variation explained by SLR

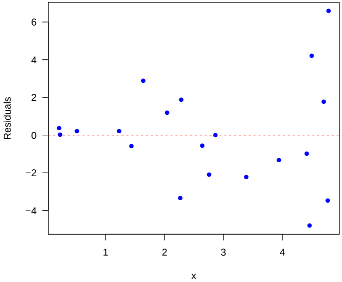
$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

- We can show $r^2 = R^2$:

$$\begin{aligned} r^2 &= \left(\hat{\beta}_{1,\text{LS}} \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}}\right)^2 \\ &= \frac{\hat{\beta}_{1,\text{LS}}^2 \sum_{i=1}^n (X_i - \bar{X})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} \\ &= \frac{\text{SSR}}{\text{SST}} \\ &= R^2 \end{aligned}$$

Residual Plot Revisited





⇒ Nonlinear relationship

- Transform X
- Nonlinear regression

⇒ Non-constant variance

- Transform Y
- Weighted least squares

Summary

In this lecture, we learned ANOVA Approach to Regression

Next time: Multiple linear regression