Lecture 4

Simple Linear Regression IV

Reading: Chapter 11

STAT 8020 Statistical Methods II
September 1, 2020

Whitney Huang
Clemson University

Agenda

1. **Analysis of Variance (ANOVA) Approach to Regression**
2. **Correlation and Coefficient of Determination**
3. **Residual Analysis: Model Diagnostics and Remedies**

ANOVA Approach to Linear Regression

Analysis of Variance (ANOVA) Approach to Regression

Partitioning Sums of Squares

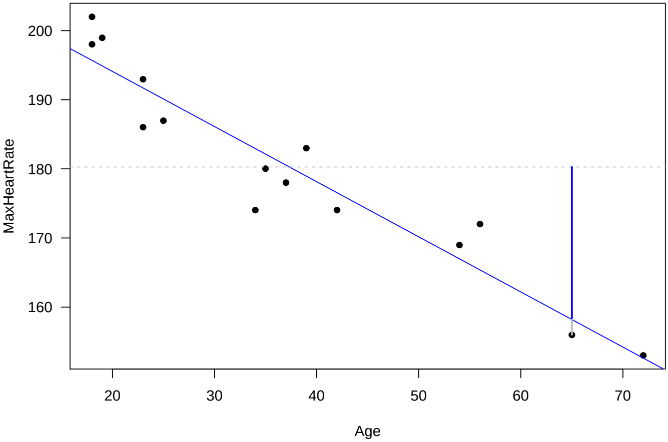
- Total sums of squares in response

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

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

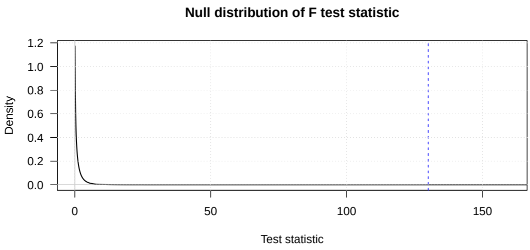
- **Goal:** To test $H_0 : \beta_1 = 0$
- Test statistics $F^* = \frac{\text{MSR}}{\text{MSE}}$
- If $\beta_1 = 0$ then F^* should be near one $\Rightarrow$ reject H_0 when F^* “large"
- We need sampling distribution of F^* under $H_0 \Rightarrow F_{1,n-2}$, where $F(d_1, d_2)$ denotes a F distribution 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
Age        1 2724.50 2724.50  130.01
Residuals 13  272.43   20.96
             Pr(>F)
Age       3.848e-08 ***
```



SLR: F-Test vs. T-test

ANOVA Table and F-Test

```
Analysis of Variance Table

Response: MaxHeartRate
          Df  Sum Sq Mean Sq
Age        1 2724.50 2724.50
Residuals 13  272.43   20.96
          F value    Pr(>F)
Age        130.01 3.848e-08
```

Parameter Estimation and T-Test

```
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
```

Correlation and Coefficient of Determination

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

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

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 $r^2 = R^2$:

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

Maximum Heart Rate vs. Age: r and R^2


```
> summary(fit)$r.squared
[1] 0.9090967
> cor(Age, MaxHeartRate)
[1] -0.9534656
```

Interpretation:

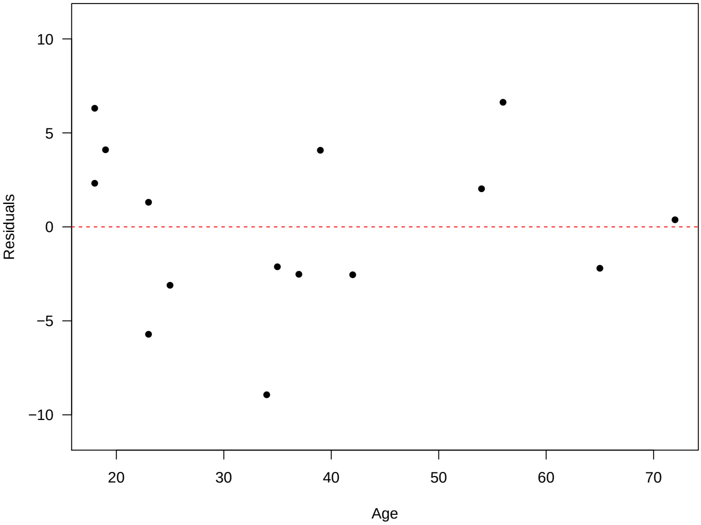
There is a strong negative linear relationship between `MaxHeartRate` and `Age`. Furthermore, $\sim 91\%$ of the variation in `MaxHeartRate` can be explained by `Age`.

Residual Analysis: Model Diagnostics and Remedies

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

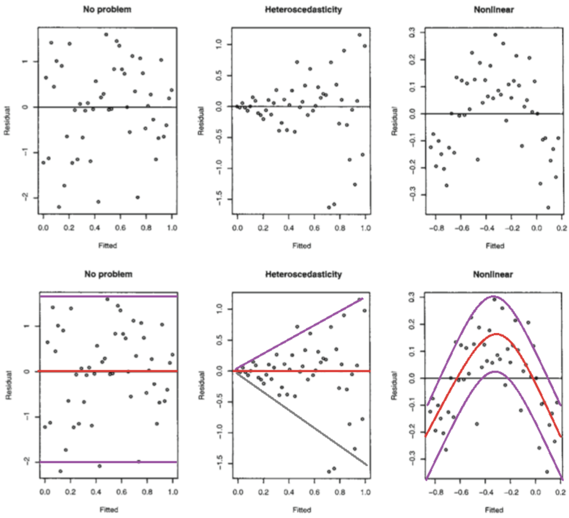
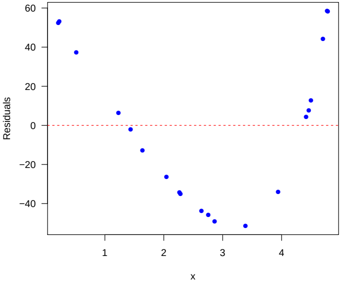


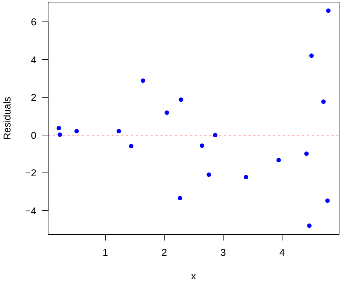
Figure: Figure courtesy of Faraway's Linear Models with R (2005, p. 59).

Model Diagnostics and Remedies



⇒ Nonlinear relationship

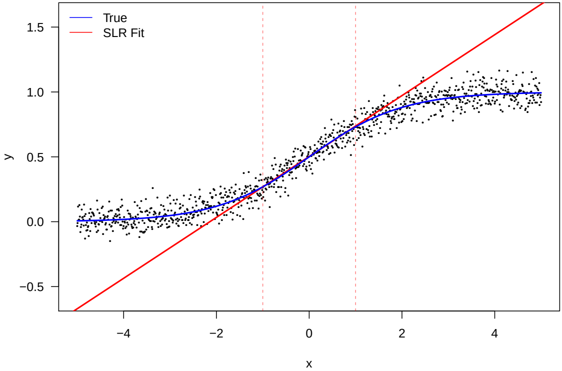
- Transform X
- Nonlinear regression



⇒ Non-constant variance

- Transform Y
- Weighted least squares

Extrapolation in SLR



Extrapolation beyond the range of the given data can lead to seriously biased estimates if the **assumed relationship does not hold the region of extrapolation**

Summary of SLR

- **Model:** $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- **Estimation:** Use the method of least squares to estimate the parameters
- **Inference**
  - **Hypothesis Testing**
  - **Confidence/prediction Intervals**
  - **ANOVA**
- **Model Diagnostics and Remedies**