Lecture 23

Simple Linear Regression II

Readings: IntroStat Chapter 11; OpenIntro Chapter 8

STAT 8010 Statistical Methods I

June 19, 2023

Whitney Huang

Clemson University

Agenda

1. Confidence/Prediction Intervals
2. Hypothesis Testing
3. Analysis of Variance (ANOVA) Approach to Regression

Recap: Simple Linear Regression

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:**
$\operatorname{argmin}_{\beta_0,\beta_1} \sum_{i=1}^n (Y_i - (\beta_0 + \beta_1 X_i))^2 \Rightarrow$
  - $\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$
  - $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$
  - $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$

- **Residuals**: $e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Recap: Residual Analysis

- **Residual Analysis:** To check the appropriateness of SLR model
  - Is the regression function linear?
  - Do ε_i's have constant variance σ^2?
  - Are ε_i's indepdent to each other?

We plot residuals e_i's against X_i's (or $\hat{Y}_i$'s) to assess these aspects

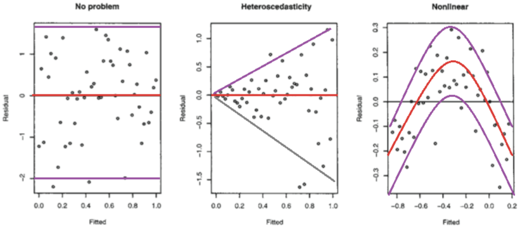
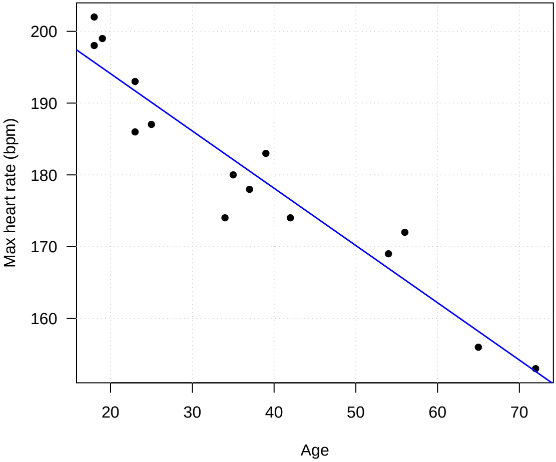


Figure: Figure courtesy of Faraway's Linear Models with R (2005, p. 59).

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:

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

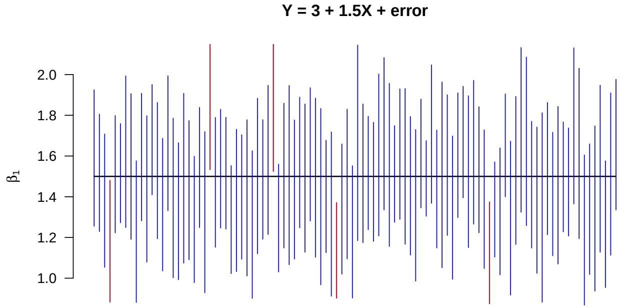
  where α is the **confidence level** and $t_{\alpha/2,n-2}$ denotes the $1-\alpha/2$ percentile of a student's t distribution with $n-2$ degrees of freedom

- Similarly, we can construct CIs for β_0:

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

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)



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:

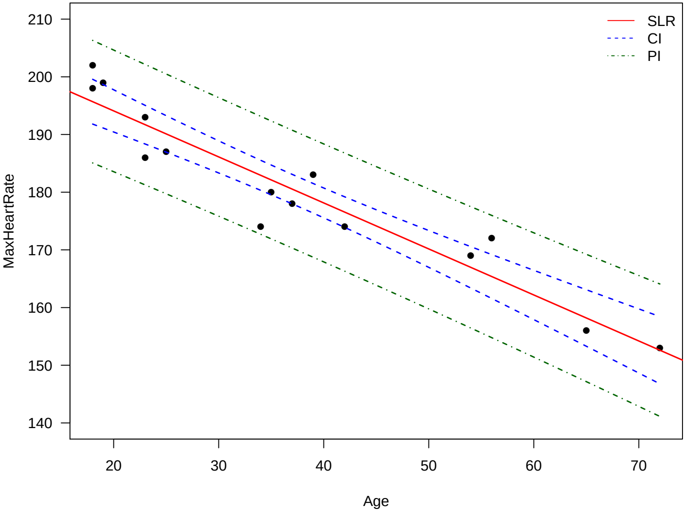
$$\left[\hat{Y}_h - t_{\alpha/2,n-2}\hat{\sigma}_{\hat{Y}_h}, \hat{Y}_h + t_{\alpha/2,n-2}\hat{\sigma}_{\hat{Y}_h}\right]$$

- **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 $Y_{\text{h(new)}}$

Confidence Intervals vs. Prediction Intervals



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 `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 $\alpha = .05$ 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**: $t^* = \frac{\hat{\beta}_0 - 0}{\hat{\sigma}_{\beta_0}} = \frac{210.0485}{2.86694} = 73.27$

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

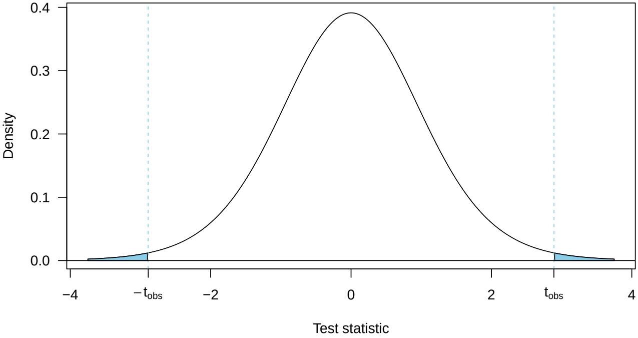
4. Compare to α and draw conclusion:

> 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 $t^* \sim t_{df=13}$

Analysis of Variance (ANOVA) Approach to Regression

Partitioning Sums of Squares

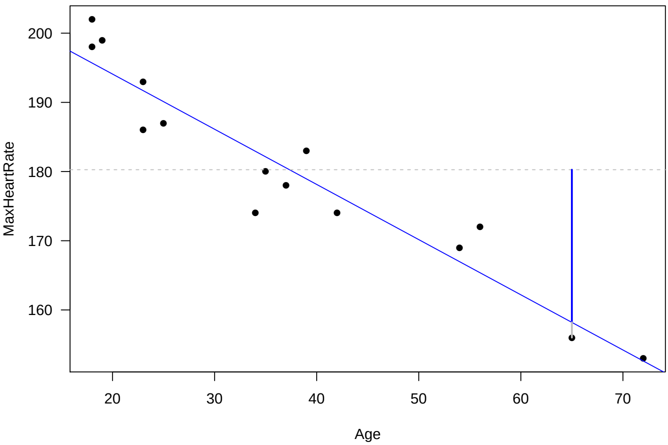
- Total sums of squares in response

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

- 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

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

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

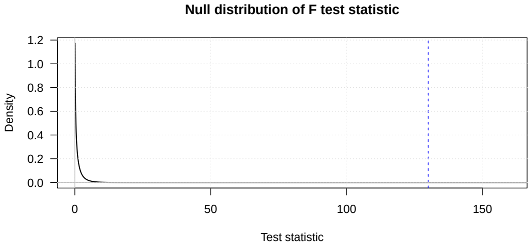
- **Goal:** To test $H_0 : \beta_1 = 0$
- Test statistics $F^* = \frac{\text{MSR}}{\text{MSE}}$
- If $\beta_1 = 0$ then F^* should be near one $\Rightarrow$ reject H_0 when F^* “large"
- We need sampling distribution of F^* under $H_0 \Rightarrow F_{1,n-2}$, where F_{d_1,d_2} denotes a F distribution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

Summary

In this lecture, we learned

- **Normal Error Regression Model** and statistical inference for β_0 and β_1
- **Confidence/Prediction Intervals** & **Hypothesis Testing**
- **ANOVA Approach to Regression**

Next time we will talk about

1. Correlation (r) & Coefficient of Determination (R^2)
2. Advanced topics in Regression Analysis