Lecture 39

Simple Linear Regression: ANOVA & Coefficient of Determination

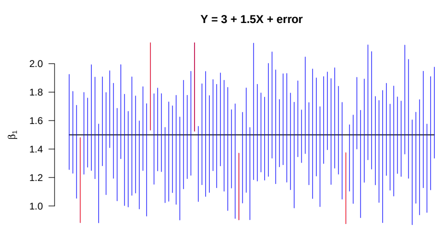
STAT 8010 Statistical Methods I
December 4, 2019

Whitney Huang
Clemson University

Notes

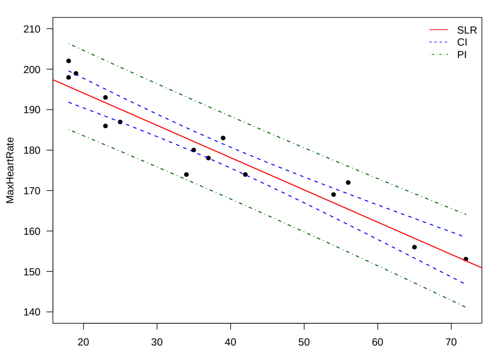
Understanding Confidence Intervals

- 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)



Notes

Confidence Intervals vs. Prediction Intervals



Notes

Analysis of Variance (ANOVA) Approach to Regression

Partitioning Sums of Squares

- Total sums of squares in response

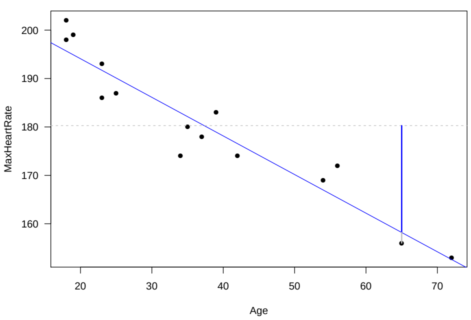
$$\text{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

- We can rewrite SST as

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

Notes

Partitioning Total Sums of Squares



Notes

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 $\text{SST}/(n-1)$ and represents an unbiased estimate of σ^2 under the model (1).

Notes

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.,
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (2)$$
- "Larg" MSR = SSR/1 suggests a linear trend, because
$$\mathrm{E}[MSE] = \sigma^2 + \beta_1^2 \sum_{i=1}^{n}(X_i - \bar{X})^2$$

Notes

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**

Notes

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

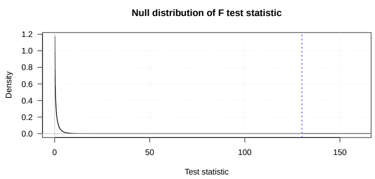
Notes

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 ***
```



Notes

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
```


Notes

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

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

Notes

Coefficient of Determination R^2

- Defined as the proportion of total variation explained by SLR

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

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

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

Notes

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

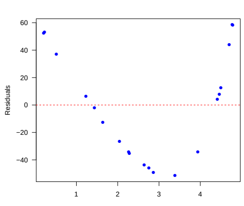
```
> summary(fit)$r.squared
[1] 0.9090967
> cor(Age, MaxHeartRate)
[1] -0.9534656
```

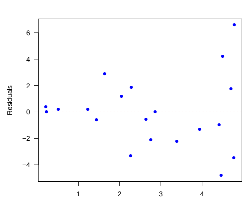
Interpretation:

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

Notes

Residual Plot Revisited





⇒ Nonlinear relationship

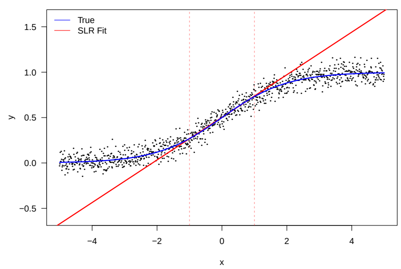
- Transform X
- Nonlinear regression

⇒ Non-constant variance

- Transform Y
- Weighted least squares

Notes

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**

Notes

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**

Simple Linear Regression: ANOVA & Coefficient of Determination
CLEMSON UNIVERSITY
Review of Last Class
Analysis of Variance (ANOVA) Approach to Regression

Notes

Summary

In this lecture, we learned ANOVA Approach to Regression and Coefficient of Determination

Next time: Review

Simple Linear Regression: ANOVA & Coefficient of Determination
CLEMSON UNIVERSITY
Review of Last Class
Analysis of Variance (ANOVA) Approach to Regression

Notes