Chapter 8

MLR Model Evaluation II

Chanwoo Yoo, Division of Advanced Engineering,
Korea National Open University

Contents

1. Sequential Sums of Squares

1. Sequential Sums of Squares

- The numerator of the general linear F-statistic — that is, SSE(R)-SSE(F) — is what is referred to as a "sequential sum of squares" or "extra sum of squares."
  - reduction in the error sum of squares (SSE) when one or more predictor variables are added to the model.
  - increase in the regression sum of squares (SSR) when one or more predictor variables are added to the model.

1. Sequential Sums of Squares

- A sequential sum of squares quantifies how much variability we explain (increase in regression sum of squares) or alternatively how much error we reduce (reduction in the error sum of squares).

2. Notation

- $SSE(x_1)$ denotes the error sum of squares when x_1 is the only predictor in the model.
- $SSR(x_1, x_2)$ denotes the regression sum of squares when x_1 and x_2 are both in the model.

2. Notation

- $SSR(x_2|x_1)$ denotes the sequential sum of squares obtained by adding x_2 to a model already containing only the predictor x_1.
- The vertical bar "|" is read as "given" — that is, " | " is read as " given ." In general, the variables appearing to the right of the bar "|" are the predictors in the original model, and the variables appearing to the left of the bar "|" are the predictors newly added to the model.

2. Notation

- $SSR(x_2|x_1) = SSE(x_1) - SSE(x_1, x_2)$
- $SSR(x_2, x_3|x_1) = SSE(x_1) - SSE(x_1, x_2, x_3)$

3. Example

```
> anova(model.1)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Area       1 0.62492 0.62492 32.1115 4.504e-06 ***
X2         1 0.31453 0.31453 16.1622  0.000398 ***
X3         1 0.01981 0.01981  1.0181  0.321602    
Residuals 28 0.54491 0.01946                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(x_1) = 0.62492$

$SSR(x_2|x_1) = 0.31453$

$SSR(x_3|x_1, x_2) = 0.01981$

4. Order Matters

```
coolhearts <- read.table("coolhearts.txt", header=T)
attach(coolhearts)

model.2 <- lm(Inf. ~ X2 + X3 + Area)
summary(model.2)
anova(model.2)
```

4. Order Matters

```
> anova(model.2)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)
X2         1 0.29994 0.29994 15.4125 0.0005124 ***
X3         1 0.02191 0.02191  1.1258 0.2977463
Area       1 0.63742 0.63742 32.7536 3.865e-06 ***
Residuals 28 0.54491 0.01946
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(x_2) = 0.29994$

$SSR(x_3|x_2) = 0.02191$

$SSR(x_1|x_2, x_3) = 0.63742$

2. The Hypothesis Tests for the Slopes

1. Research Question I

- Is a regression model containing at least one predictor useful in predicting the size of the infarct?
  - $H_0: \beta_1 = \beta_2 = \beta_3 = 0$
  - H_A: At least one $\beta_i \neq 0$ (for $i = 1,2,3$)

1. Research Question I

- The reduced model: $y_i = \beta_0 + \epsilon_i, df_R = n - 1$
- The full model: $y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i, df_F = n - 4$
- Overall F−test
  - $F^* = \left(\frac{SSE(R)-SSE(F)}{df_R - df_F}\right) \div \left(\frac{SSE(F)}{df_F}\right) = \frac{SSR}{3} \div \frac{SSE}{n-4} = \frac{MSR}{MSE}$

1. Research Question I

```
> anova(model.1)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)    
Area       1 0.62492 0.62492 32.1115 4.504e-06 ***
X2         1 0.31453 0.31453 16.1622  0.000398 ***
X3         1 0.01981 0.01981  1.0181  0.321602    
Residuals 28 0.54491 0.01946                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

df_F SSE MSE

$SSR = 0.62492 + 0.31453 + 0.01981$
$= 0.95926$

$MSR = 0.95926/3 = 0.31975$

$MSE = 0.54491/28 = 0.01946$

1. Research Question I

```
> summary(model.1)


...
Residual standard error: 0.1395 on 28 degrees of freedom
Multiple R-squared:  0.6377,  Adjusted R-squared:  0.5989
F-statistic: 16.43 on 3 and 28 DF,  p-value: 2.363e-06
```

$$F^* = \frac{MSR}{MSE} = \frac{0.31975}{0.01946} = 16.43$$

1. Research Question I

- There is sufficient evidence ($F = 16.43$, $P < 0.001$) to conclude that at least one of the slope parameters is not equal to 0.
- In general, to test that all of the slope parameters in a multiple linear regression model are 0, we use the overall F-test.

2. Research Question II

- Is the size of the infarct significantly (linearly) related to the area of the region at risk?
  - $H_0: \beta_1 = 0$
  - $H_A: \beta_1 \neq 0$

2. Research Question II

- The reduced model: $y_i = (\beta_0 + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i, df_R = n - 3$
- The full model: $y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i, df_F = n - 4$
- $F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F}\right) \div \left(\frac{SSE(F)}{df_F}\right) = \frac{SSR(x_1|x_2, x_3)}{1} \div \frac{SSE(x_1, x_2, x_3)}{n-4} =$

  $\frac{MSR(x_1|x_2, x_3)}{MSE(x_1, x_2, x_3)}$

2. Research Question II

```
coolhearts <- read.table("coolhearts.txt", header=T)
attach(coolhearts)

model.2 <- lm(Inf. ~ X2 + X3 + Area)
summary(model.2)
anova(model.2)
```

2. Research Question II

```
> anova(model.2)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)    
X2         1 0.29994 0.29994 15.4125 0.0005124 ***
X3         1 0.02191 0.02191  1.1258 0.2977463    
Area       1 0.63742 0.63742 32.7536 3.865e-06 ***
Residuals 28 0.54491 0.01946                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$SSR(x_1|x_2, x_3) = 0.63742$

$MSR(x_1|x_2, x_3) = 0.63742$

$MSE(x_1, x_2, x_3)$

$= 0.54491/28 = 0.01946$

$MSE(x_1, x_2, x_3)$

df_F $SSE(x_1, x_2, x_3)$

2. Research Question II

- $MSR(x_1|x_2,x_3) = 0.63742$
- $MSE = 0.01946$
- $F^* = \frac{MSR(x_1|x_2,x_3)}{MSE(x_1,x_2,x_3)} = \frac{0.63742}{0.01946} = 32.7554$
    - 1 numerator degree of freedom and 28 denominator degree of freedom

2. Research Question II

```
> pf(32.7554, 1, 28, lower.tail = FALSE)
[1] 3.863795e-06
```

- There is sufficient evidence ($F = 32.7554$, $P < 0.001$) to conclude that the size of the infarct is significantly related to the size of the area at risk after the other predictors X2 and X3 have been taken into account.

2. Research Question II

```
> anova(model.2)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)    
X2         1 0.29994 0.29994 15.4125 0.0005124 ***
X3         1 0.02191 0.02191  1.1258 0.2977463    
Area       1 0.63742 0.63742 32.7536 3.865e-06 ***
Residuals 28 0.54491 0.01946                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2. Research Question II

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.13454    0.10402  -1.293 0.206459
X2          -0.24348    0.06229  -3.909 0.000536 ***
X3          -0.06566    0.06507  -1.009 0.321602
Area         0.61265    0.10705   5.723 3.87e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$(5.723)^2 = 32.75$

$t_{(n-p)}^2 = F_{(1,n-p)}$

2. Research Question II

- We can use either the F-test or the t-test to test that only one slope parameter is 0.
- The equivalence of the t-test to the F-test has taught us something new about the t-test. The t-test is a test for the marginal significance of the predictor after the other predictors and have been taken into account. It does not test for the significance of the relationship between the response y and the predictor alone.

3. Research Question III

- Is the size of the infarct area significantly (linearly) related to the type of treatment after controlling for the size of the region at risk for infarction?
  - $H_0: \beta_2 = \beta_3 = 0$
  - H_A: At least one $\beta_i \neq 0$ (for $i = 2,3$)

3. Research Question III

- The reduced model: $y_i = (\beta_0 + \beta_1 x_{i1}) + \epsilon_i, df_R = n - 2$
- The full model: $y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i, df_F = n - 4$
- $F^* = \left(\frac{SSE(R) - SSE(F)}{df_R - df_F}\right) \div \left(\frac{SSE(F)}{df_F}\right) = \frac{SSR(x_2, x_3 | x_1)}{2} \div \frac{SSE(x_1, x_2, x_3)}{n-4} =$

  $\frac{MSR(x_2, x_3 | x_1)}{MSE(x_1, x_2, x_3)}$

3. Research Question III

```
coolhearts <- read.table("coolhearts.txt", header=T)
attach(coolhearts)

model.3 <- lm(Inf. ~ Area)
anova(model.3)
```

3. Research Question III

```
> anova(model.3)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)
Area       1 0.62492 0.62492  21.322 6.844e-05 ***
Residuals 30 0.87926 0.02931
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSE(R)$

3. Research Question III

```
> anova(model.1)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)
Area       1 0.62492 0.62492 32.1115 4.504e-06 ***
X2         1 0.31453 0.31453 16.1622  0.000398 ***
X3         1 0.01981 0.01981  1.0181  0.321602
Residuals 28 0.54491 0.01946
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSE(F)$ $MSE(F)$

$$SSE(R) - SSE(F)$$
$$= SSE(x_1) - SSE(x_1, x_2, x_3)$$
$$= 0.87926 - 0.54491 = 0.33435$$
$$= 0.31453 + 0.01981$$

3. Research Question III

- $SSE(R) - SSE(F) = 0.33435$
- $MSE(F) = 0.01946$
- $F^* = \left(\frac{SSE(R)-SSE(F)}{df_R - df_F}\right) \div \left(\frac{SSE(F)}{df_F}\right) = \frac{SSR(x_2,x_3|x_1)}{2} \div \frac{SSE(x_1,x_2,x_3)}{n-4} =$

  $\frac{MSR(x_2,x_3|x_1)}{MSE(x_1,x_2,x_3)} = \frac{0.33435}{2} \div 0.01946 = 8.59$

  - 2 numerator degree of freedom and 28 denominator degree of freedom

3. Research Question III

```
> pf(8.59, 2, 28, lower.tail = FALSE)
[1] 0.001233006
```

- There is sufficient evidence ($F = 8.59$, $P = 0.0012$) to conclude that the type of cooling is significantly related to the extent of damage that occurs — after taking into account the size of the region at risk.

3. Summary

1. Summary

- Hypothesis test for testing that all of the slope parameters are 0.
- Hypothesis test for testing that one slope parameter is 0.
- Hypothesis test for testing that a subset — more than one, but not all — of the slope parameters are 0.

1. Summary

- Hypothesis test for testing that all of the slope parameters are 0.
  - Overall F-test: F-statistic and associated p-value in model summary

```
> summary(model.1)

...
Residual standard error: 0.1395 on 28 degrees of freedom
Multiple R-squared:  0.6377,   Adjusted R-squared:  0.5989
F-statistic: 16.43 on 3 and 28 DF,  p-value: 2.363e-06
```

1. Summary

- Hypothesis test for testing that one slope parameter is 0.
  - General Linear F−test or t−test

1. Summary

```
> anova(model.2)
Analysis of Variance Table

Response: Inf.
          Df  Sum Sq Mean Sq F value    Pr(>F)    
X2         1 0.29994 0.29994 15.4125 0.0005124 ***
X3         1 0.02191 0.02191  1.1258 0.2977463    
Area       1 0.63742 0.63742 32.7536 3.865e-06 ***
Residuals 28 0.54491 0.01946                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. Summary

```
Coefficients:
            Estimate Std. Error t value Pr(>¦t¦)
(Intercept) -0.13454    0.10402  -1.293 0.206459
X2          -0.24348    0.06229  -3.909 0.000536 ***
X3          -0.06566    0.06507  -1.009 0.321602
Area         0.61265    0.10705   5.723 3.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

$(5.723)^2 = 32.75$

$t^2_{(n-p)} = F_{(1,n-p)}$

1. Summary

- Hypothesis test for testing that a subset — more than one, but not all — of the slope parameters are 0.
  - General Linear F−test

Next

Chapter 9
MLR Estimation, Prediction & Model Assumptions