STAC67: Regression Analysis

Lecture 13

Sohee Kang

Mar. 2, 2021

Chapter 7: Multiple Regresssion II

Testing a Subset of Coefficients

- We may want to test if some but not all the coefficents are 0.
- We define full model to be:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p + \epsilon_i$$

- Suppose the null hypothesis we want to test is:

$$H_0 : \beta_{k+1} = \beta_{k+2} = \ldots = \beta_p = 0$$
$$H_a : \text{at least one } \beta_i \neq 0$$

- Then we can define the reduced model to be:

$$Y_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

- From the full model, we get SSE (SSE_F) and MSE (MSE_F) with the degrees of freedom: $n-p'$
- From the reduced model, we get (SSE_R) with degrees of freedom: $n-k'$

Testing a subset of coefficients

$$F = \frac{\frac{SSE_R - SSE_F}{df_R - df_F}}{MSE_F} = \frac{\frac{SSR_F - SSR_R}{p-k}}{MSE_F} \sim F(p-k, n-p')$$

$p-k$

Further Decomposition of Sum of Squares

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \beta_p X_p + \epsilon_i$$

- Series of submodels (or reduced models)

(X_1) : $\quad Y = \beta_0 + \beta_1 X_1 + \varepsilon$

(X_1, X_2) : $\quad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

$\vdots \qquad \vdots$

$(X_1, X_2, \ldots, X_p)$: $\quad Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

Decomposition of sum of squares

- For each model, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} \ldots + \beta_p X_{ip} + \epsilon_i$
- We can calculate its SST : $\sum (Y_i - \bar{Y})^2 = Y'Y - \frac{1}{n} Y'JY = Y'(I - \frac{1}{n}J)Y$

which is the same for all models

and SSR$(X_1, \ldots, X_p)$ and SSE$(X_1, \ldots, X_p)$

Models	SSR	SSE	Extra Sum of Squares
X_1	$SSR(X_1)$	$SSE(X_1)$	
X_1, X_2	$SSR(X_1, X_2)$	$SSE(X_1, X_2)$	$SSR(X_2 \mid X_1) = SSR(X_1, X_2) - SSR(X_1) = SSE(X_1) - SSE(X_1, X_2)$
$\vdots$			
$(X_1, X_2 \cdots X_p)$	$SSR(X_1, X_2 \cdots X_p)$	$SSE(X_1, X_2, \ldots X_p)$	$SSR(X_p \mid X_1, \ldots X_{p-1}) = SSR(X_1, \cdots X_p) - SSR(X_1, \ldots, X_{p-1})$

Decomposition of sum of squares

- For any model:

$$SST = SSE(X_1, \ldots, X_p) + SSR(X_1, \ldots, X_p)$$

$$SSR(X_1, \ldots, X_p) = SSR(X_1) + SSR(X_2 | X_1) + \ldots + SSR(X_p | X_1, \ldots, X_{p-1})$$

$$SST = SSE(X_1, \ldots, X_p) + SSR(X_1, \ldots, X_p) \leftarrow \text{as above}$$

- Decompostion of degrees of freedom

Source	DF
$SSR(X_1)$	1
$SSR(X_2 \mid X_1)$	1
$\vdots$	$\vdots$
$SSR(X_p \mid X_1, \ldots, X_{p-1})$	1
total sum	p

Interpretation of SSE and SSR

- $SSE(X_1)$: variation of Y unexplained by X_1
- $SSR(X_1)$: Variation of Y explained by X_1
- $SSE(X_1, X_2)$: variation of Y unexplained by X_1 and X_2
- $SSR(X_1, X_2)$: Variation of Y explained by X_1 and X_2
- $SSR(X_2|X_1)$: extra sum of squares, additional variation explained by introducing X_2.

Example: BodyFat Data

- The data consists of 20 females whose age are between 25 and 30 years old.
- Variables in the data set are:
  - y = amount of body fat (percentage) $\quad x_1 =$ triceps skinfold thickness,
  - $x_2 =$ thigh circumference $\quad x_3 =$ midarm circumference

```
Data = read.table("bodyfat.txt")
names(Data) = c("X1", "X2", "X3", "Y")
Data[1:3, ]
```

```
##     X1   X2   X3    Y
## 1 19.5 43.1 29.1 11.9
## 2 24.7 49.8 28.2 22.8
## 3 30.7 51.9 37.0 18.7
```

```
fit = lm(Y~X1 + X2 +X3, data=Data)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7263 -1.6111  0.3923  1.4656  4.1277
##
## Coefficients:
```

BodyFat Example

- Model 1: regression of Y on X1: $\hat{Y} = -1.4296 + 0.8572 X_1$

```
anova(lm(Y~X1, data=Data))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27  44.305 3.024e-06 ***
## Residuals 18 143.12    7.95
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 2: regression of Y on X2:

```
anova(lm(Y~X2, data=Data))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X2         1 381.97  381.97  60.617 3.6e-07 ***
## Residuals 18 113.42    6.30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

BodyFat Example

- Model 3: regression of Y on X1 and X2:

```
anova(lm(Y~X1+X2, data=Data))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 54.4661 1.075e-06 ***
## X2         1  33.17   33.17  5.1284    0.0369 *
## Residuals 17 109.95    6.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Model 4: regression of Y on X1, X2, X3:

```
anova(lm(Y~X1+X2+X3, data=Data))
```

SSR(X1)

SSR(X2|X1)

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X1         1 352.27  352.27 57.2768 1.131e-06 ***
## X2         1  33.17   33.17  5.3931   0.03373 *
## X3         1  11.55   11.55  1.8773   0.18956
## Residuals 16  98.40    6.15
## ---
```

SSE(X1, ... Xp)

BodyFat Example

- Test for regression coefficients

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

$$H_0 : \beta_3 = 0 \quad vs \quad H_a : \beta_3 \neq 0$$

- Full Model: $Y = B_0 + B_1 X_1 + B_2 X_2 + B_3 X_3$
- Reduced Model: $Y = B_0 + B_1 X_1 + B_2 X_2$

$$F^* = \frac{\frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{(n-3) - (n-4)}}{6.15} = \frac{\frac{109.95 - 98.4}{1}}{6.15} = 1.88$$

$$F(\alpha = 0.01, 1, 16) = 8.53$$

Since $F^* < F_{val}$, we cannot reject H_0.

$$F^* = \frac{\frac{SSR(X_1, X_2, X_3) - SSR(X_1, X_2)}{1}}{MSE_F}$$

$$= SSR(X_3 \mid X_2, X_2) + SSR(X_2 \mid X_1) + SSR(X_1) - \left(SSR(X_2 \mid X_1) + SSR(X_1)\right)$$

$$= \frac{SSR(X_3 \mid X_2, X_1)}{MSE_F}$$

$$= \frac{11.55}{6.15}$$