Chapter 2 Multiple Regression
(Part 3)

1 Further decomposition of sums of squares

Consider general model

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, ..., n$$

and a series of sub-models (or reduced models)

$$\begin{aligned}
(X_1): &\qquad Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i, \quad i = 1, ..., n \\
(X_1, X_2): &\qquad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i, \quad i = 1, ..., n \\
... & \\
(X_1, X_2, ..., X_p): &\qquad Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, ..., n
\end{aligned}$$

For each model, say $Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_p X_{ik} + \varepsilon_i$, we can calculate its SST (which is the same for all models) and

$$SSR(X_1, ..., X_k), \quad SSE(X_1, ..., X_k).$$

We have the sum of squares of regressions as follows

models	SSR	SSE	extra SS
(X_1)	SSR(X_1)	SSE(X_1)	—
(X_1, X_2)	SSR(X_1, X_2)	SSE(X_1, X_2)	SSR($X_2 \vert X_1$) = SSR(X_1, X_2)-SSR(X_1) = SSE(X_1)-SSE(X_1, X_2)
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(X_1, ..., X_p)$	SSR($X_1, ..., X_p$)	SSE($X_1, ..., X_p$)	SSR($X_p \vert X_1, ..., X_{p-1}$) = SSR($X_1, ..., X_p$) -SSR($X_1, ..., X_{p-1}$) = SSE($X_1, ..., X_{p-1}$)-SSE($X_1, ..., X_p$)

It is easy to see that

- for any model

$$SST = SSE(X_1, X_2, ..., X_k) + SSR(X_1, X_2, ..., X_k), \qquad k = 1, 2, ..., p$$

-

$$SSR(X_1, ..., X_k) = SSR(X_1)+SSR(X_2|X_1)+...+SSR(X_k|X_1, ..., X_{k-1}) \qquad k = 1, 2, ..., p$$

-

$$SST = SSE(X_1, ..., X_p)+SSR(X_1)+SSR(X_1|X_2)+....+SSR(X_k|X_1, ..., X_{k-1}), \qquad k = 1, 2, ..., p$$

- Degree of freedom (D.F.)

	source	D.F.
	SSR(X_1):	1
	SSR($X_2|X_1$):	1
	$\vdots$	
	SSR($X_p|X_1, ..., X_{p-1}$):	1
Total	SSR($X_1, ..., X_p$):	p

In multiple regression, the ANOVA table is (sometimes)

source of variateion	SS	D.F.	MS	F-value	$P - value$
X_1	SSR(X_1)	1	MSR(X_1)	MSR(X_1)/MSE	
$X_2|X_1$	SSR($X_2|X_1$)	1	MSR($X_2|X_1$)	MSR($X_2|X_1$)/MSE	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$X_p|(X_1, ..., X_{p-1})$	SSR($X_p|(X_1, ..., X_{p-1})$)	1	MSR($X_p|(X_1, ..., X_{p-1})$)	$\frac{MSR(X_p|(X_1,...,X_{p-1}))}{MSE}$	
Error	SSE($X_1, ..., X_p$)	n-p-1	MSE = $\frac{SSE(X_1,...,X_p)}{n-p-1}$		

where $P - value$ is the probability $P(F(1, n - p - 1) > F$-value)

1.1 Interpretation of SSE and SSR

- SSE(X_1) — SSE of model 1: variation of Y unexplained by X_1

  SSR(X_1) — SSR of model 1: variation of Y explained by X_1

- SSE(X_1, X_2) — SSE of model 3: variation of Y unexplained by X_1 and X_2

  SSR(X_1, X_2) — SSR of model 2: variation of Y explained by X_1 and X_2

  SSR($X_2|X_1$)= SSR(X_1, X_2)-SSR(X_1) — additional/extra sum of square (extra variation explained) due to introducing X_2 after X_1 is introduced.

- SSE(X_1, X_2, X_3) — SSE of model 4: variation of Y unexplained by X_1, X_2 and X_3

  SSR(X_1, X_2, X_3) — SSR of model 4: variation of Y explained by X_1, X_2 and X_3

  SSR($X_3|X_1, X_2$)= SSR(X_1, X_2, X_3)-SSR(X_1, X_2) — additional/extra sum of square (extra variation explained) due to introducing X_3 after X_1 and X_2 are introduced.

  ......

- Therefore, SSR($X_{k+1}|X_1, X_2, ..., X_k$) can be used to check whether we need to introduce more variables after $X_1, ..., X_k$ are introduced.

1.2 model testing and extension

Testing hypothesis about the whole model (see lecture notes Part 2 of Chapter 2.)

Testing hypothesis about parts of the model *We use one example to explain the idea.* Consider two models

$$\text{Full model: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_i$$

and

$$\text{Reduced model: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$$

Then the extra sum of squares (variation) explained by adding/introducing X_4, X_5 to the "reduced model" is

$$\begin{aligned} SSR(X_4, X_5|X_1, X_2, X_3) &= SSR(X_1, X_2, X_3, X_4, X_5) - SSR(X_1, X_2, X_3) \\ &= SSE(X_1, X_2, X_3) - SSE(X_1, X_2, X_3, X_4, X_5) \end{aligned}$$

with degree of freedom:

$$\begin{aligned} &\text{DF of SSE}(X_1, X_2, X_3) - \text{DF of SSE}(X_1, X_2, X_3, X_4, X_5) \\ &= (n-3-1) - (n-5-1) = 2. \end{aligned}$$

where 2 is the difference of numbers of variables in the tow models. We write

$$df(F) = \text{DF of SSE}(X_1, X_2, X_3, X_4, X_5), \quad df(R) = \text{DF of SSE}(X_1, X_2, X_3),$$

If the extra sum of squares (extra variation explained) is "big", it is necessary need to introduce X_4, X_5. Otherwise, it is not necessary. Consider hypothesis

$$H_0 : \beta_4 = \beta_5 = 0, \quad v.s. \quad H_1 : \text{not all of them are } 0$$

We consider the F-statstic

$$F = \frac{SSR(X_4, X_5|X_1, X_2, X_3)/(df(R) - df(F))}{SSE(F)/df(F)}$$

Under H_0,

$$F \sim F(df(R) - df(F), df(F))$$

For significant level α and calculated F-value, denoted by F^*,

- If $F^* > F(1-\alpha, df(R) - df(F), df(F))$, we reject H_0.
- If $F^* \leq F(1-\alpha, df(R) - df(F), df(F))$, we accept H_0.

2 An example: Body fat

- Response variable: Y - amount of body fat
- X_1: triceps skinfold thickness
- X_2: thigh circumference
- X_3: midarm circumference
- **Data**

20 healthy females 25-34 years old

individual	X_1	X_2	X_3	Y
1	19.5	43.1	29.1	11.9
2	24.7	49.8	28.2	22.8
⋮	⋮	⋮	⋮	⋮
19	22.7	48.2	27.1	14.8
20	25.2	51.0	27.5	21.1

- models and ANOVA tables

Model 1: regression of Y on X_1: $\hat{Y} = -1.496 + 0.8572X_1$

Source of variation	SS	df
Regression	352.27	1
Error	143.12	18
Total	495.39	19

Model 2: regression of Y on X_2: $\hat{Y} = -23.634 + 0.8565X_2$

Source of variation	SS	df
Regression	381.97	1
Error	113.42	18
Total	495.39	19

Model 3: regression of Y on X_1 and X_2: $\hat{Y} = -19.174 + 0.2224X_1 + 0.6594X_2$

Source of variation	SS	df
Regression	385.44	2
Error	109.95	17
Total	495.39	19

Model 4: regression of Y on X_1, X_2 and X_3:

$\hat{Y} = 117.08 + 4.334X_1 - 2.857X_2 - 2.186X_3$

Source of variation	SS	df
Regression	396.98	3
Error	98.41	16
Total	495.39	19

- Extra sums of squares
  - the additional/extra sum of square (extra variation explained) by adding X_2 to model 1:

$$\begin{aligned} SSR(X_2|X_1) &= SSR(X_1, X_2) - SSR(X_1) = SSE(X_1) - SSE(X_1, X_2) \\ &= 143.12 - 109.95 = 33.17 \end{aligned}$$

  - the additional/extra sum of square (extra variation explained) by adding X_3 to model 3:

$$\begin{aligned} SSR(X_3|X_1, X_2) &= SSR(X_1, X_2, X_3) - SSR(X_1, X_2) \\ &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \\ &= 109.95 - 98.41 = 11.54 \end{aligned}$$

2.1 ANOVA for the body fat example

```
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

---
Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1
```

2.2 Tests for regression coefficients

- Assume $Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$

  Test $H_0 : \beta_3 = 0$ versus $H_a : \beta_3 \neq 0$

- General linear test approach:

  Full model (under H_a): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

  Reduced model (under H_0): $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

  Let $df(F)$ be the degree of freedom of SSE for the full model

  Let $df(R)$ be the degree of freedom of SSE for the reduced model

$$
\begin{aligned}
F^* &= \frac{(SSE(R) - SSE(F))/(df(R) - df(F))}{SSE(F)/df(F)} \\
&= \frac{(SSE(X_1, X_2) - SSE(X_1, X_2, X_3))/1}{SSE(X_1, X_2, X_3)/(20-4)} \\
&= \frac{SSE(X_3|X_1, X_2)}{SSE(X_1, X_2, X_3)/16} \\
&= \frac{11.54/1}{98.41/16}
\end{aligned}
$$

  $1.88 \leq F(0.99, 1, 16) = 8.53$, we accept H_0 and the reduced model with $\alpha = 0.01$

2.3 Tests for regression coefficients

- Assume $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

  Test $H_0 : \beta_2 = \beta_3 = 0$ versus $H_a : \beta_2 \neq 0$ or $\beta_3 \neq 0$

- General linear test approach:

  Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

  Reduced model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

$$
\begin{aligned}
F^* &= \frac{(SSE(R) - SSE(F))/(df(R) - df(F))}{SSE(F)/df(F)} \\
&= \frac{(SSE(X_1) - SSE(X_1, X_2, X_3))/2}{SSE(X_1, X_2, X_3)/(20-4)} \\
&= \frac{SSE(X_2, X_3|X_1)}{SSE(X_1, X_2, X_3)/16} \\
&= ((143.120 - 98.41)/2)/(98.41/16) = 3.6346 > F(0.95, 2, 16) = 3.63
\end{aligned}
$$

  So, we reject H_0, that is at least one of β_2 and β_3 are 0. Or introducing (X_2, X_3) is necessary.

2.4 Other Tests for regression coefficients

- We might want to test

$$H_0 : \beta_1 = \beta_2, \quad H_a : \beta_1 \neq \beta_2$$

  Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

  Reduced model: $Y = \beta_0 + \beta_1 (X_1 + X_2) + \beta_3 X_3 + \varepsilon$

  Test statistic

$$F = \frac{(SSE(R) - SSE(F))/(df(R) - df(F))}{SSE(F)/df(F)}$$

  with

$$df(R) - df(F) = 1$$

  and

$$df(F) = n - 4$$

  How to make conclusion?

- We might want to test

$$H_0 : \beta_1 = 3, \beta_2 = 5, \quad H_a : \text{not all equalities in } H_0 \text{ hold}$$

Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Reduced model: $Y = \beta_0 + 3X_1 + 5X_2 + \beta_3 X_3 + \varepsilon$

Test statistic

$$F = \frac{(SSE(R) - SSE(F))/(df(R) - df(F))}{SSE(F)/df(F)}$$

with

$$df(R) - df(F) = 2,$$

and

$$df(F) = n - 4$$

How to make conclusion?

2.5 Coefficient of Partial determination [advanced topics]

Recall that for the simple linear regression model, the slop coefficient is strongly related with the linear correlation coefficients. But this relationship does not hold for multiple regression model.

A **Coefficient of partial determination** measure the marginal contribution of one X variable when all the others are already included in the model.

The definition is as follows

- Given X_1 is included, the partial R^2 of X_2, denoted by $R^2_{Y2|1}$

$$R^2_{Y2|1} = \frac{SSR(X_2|X_1)}{SSE(X_1)} = \frac{SSE(X_1) - SSE(X_1, X_2)}{SSE(X_1)}$$

- Given X_1, X_2 is included, the partial R^2 of X_3, denoted by $R^2_{Y3|12}$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = \frac{SSE(X_1, X_2) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_2)}$$

- Given X_1, X_3 is included, the partial R^2 of X_2, denoted by $R^2_{Y2|13}$

$$R^2_{Y2|13} = \frac{SSR(X_2|X_1, X_3)}{SSE(X_1, X_3)} = \frac{SSE(X_1, X_3) - SSE(X_1, X_2, X_3)}{SSE(X_1, X_3)}$$

For the Body fat example

- SST=495.39, SSR(X_1)=352.27, SSR(X_2)=381.97

- coefficient of determination R^2 measures the proportion of variation explained by X

$$R^2_{Y1} = \frac{SSR(X_1)}{SST} = 0.71$$

- Coefficient of Partial determination measures the proportion explained by one additional X

$$R^2_{Y1|2} = \frac{SSR(X_1|X_2)}{SSE(X_2)} = 0.031$$

$$R^2_{Y3|12} = \frac{SSR(X_3|X_1, X_2)}{SSE(X_1, X_2)} = 0.105$$