Stat 206: Linear Models

Lecture 8

October 21, 2015

Extra Sum of Squares

$\mathcal{I}$ and $\mathcal{J}$ are two **non-overlapping** index sets.

- **Extra sum of squares (ESS)**:

$$SSR(X_{\mathcal{J}}|X_{\mathcal{I}}) := \qquad\qquad\qquad\qquad .$$

- It indicates the

- Degrees of freedom: $d.f.(SSR(X_{\mathcal{J}}|X_{\mathcal{I}})) =$.
- Mean squares: $MSR(X_{\mathcal{J}}|X_{\mathcal{I}}) :=$.

Extra Sum of Squares

$\mathcal{I}$ and $\mathcal{J}$ are two **non-overlapping** index sets.

- **Extra sum of squares (ESS)**:

$$SSR(X_{\mathcal{J}}|X_{\mathcal{I}}) := SSE(X_{\mathcal{I}}) - SSE(X_{\mathcal{I}}, X_{\mathcal{J}}).$$

- It indicates the **reduction in error sum of squares by adding $X_{\mathcal{J}}$ to the model with $X_{\mathcal{I}}$ being the X variables**.
- Degrees of freedom: $d.f.(SSR(X_{\mathcal{J}}|X_{\mathcal{I}})) = |\mathcal{J}|$.
- Mean squares: $MSR(X_{\mathcal{J}}|X_{\mathcal{I}}) := \frac{SSR(X_{\mathcal{J}}|X_{\mathcal{I}})}{d.f.(SSR(X_{\mathcal{J}}|X_{\mathcal{I}}))}$.

Notations.

- $\mathcal{I}$: an index set; $X_{\mathcal{I}} := \{X_i : i \in \mathcal{I}\}$.
  - E.g. $\mathcal{I} = \{2, 3\}$, $X_{\mathcal{I}} = \{X_2, X_3\}$.
- $SSE(X_{\mathcal{I}})$ and $SSR(X_{\mathcal{I}})$ denote the error sum of squares and regression sum of squares, respectively, under the regression model with $X_{\mathcal{I}} := \{X_i : i \in \mathcal{I}\}$ being the X variables.
  - E.g., $SSE(X_2, X_3)$ is the error sum of squares of the model with X_2 and X_3.

Some properties of ESS.

- $SSR(X_{\mathcal{J}}|X_{\mathcal{I}})$.
- Usually $SSR(X_{\mathcal{J}}|X_{\mathcal{I}})$ $SSR(X_{\mathcal{I}}|X_{\mathcal{J}})$.
- ESS is also the marginal of the regression sum of squares, i.e.,

$$SSR(X_{\mathcal{J}}|X_{\mathcal{I}}) = \qquad .$$

- SSR of a model with only one X variable may be viewed as an ESS.
  - ϕ denotes the empty set. Then $SSR(X_\phi) =$, and

$$SSR(X_1|X_\phi) = \qquad ,$$

    i.e., $SSR(X_1)$ is the of the regression sum of squares by adding X_1 into a model with only intercept but no X variable.

Some properties of ESS.

- $SSR(X_{\mathcal{J}}|X_{\mathcal{I}}) \geq 0$.
- Usually $SSR(X_{\mathcal{J}}|X_{\mathcal{I}}) \neq SSR(X_{\mathcal{I}}|X_{\mathcal{J}})$.
- ESS is also the marginal increase of the regression sum of squares, i.e.,

$$SSR(X_{\mathcal{J}}|X_{\mathcal{I}}) = SSR(X_{\mathcal{I}}, X_{\mathcal{J}}) - SSR(X_{\mathcal{I}}).$$

- SSR of a model with only one X variable may be viewed as an ESS.
  - ϕ denotes the empty set. Then $SSR(X_\phi) = 0$, and

$$SSR(X_1|X_\phi) = SSR(X_1, X_\phi) - SSR(X_\phi) = SSR(X_1),$$

    i.e., $SSR(X_1)$ is the increase of the regression sum of squares by adding X_1 into a model with only intercept but no X variable.

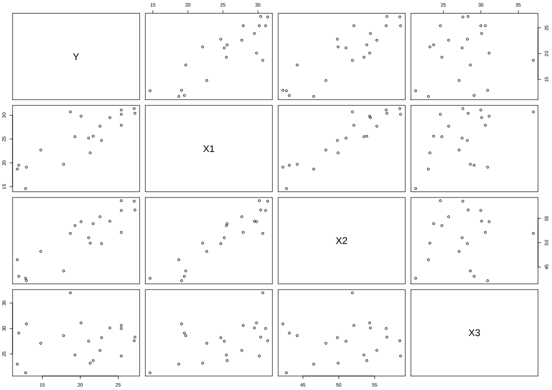
Body Fat

A researcher measured the amount of body fat (Y) of 20 healthy females 25 to 34 years old, together with three (potential) predictor variables, triceps skinfolds thickness (X_1), thigh circumference (X_2), and midarm circumference (X_3). The amount of body fat was obtained by a cumbersome and expensive procedure requiring immersion of the person in water. Thus it would be helpful if a regression model with some or all of these predictors could provide reliable estimates of body fat as these predictors are easy to measure.

A snapshot of the data.

case	X1	X2	X3	Y
Triceps	Thigh	MidArm	BodyFat	
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
4	29.8	54.3	31.1	20.1
5	19.1	42.2	30.9	12.9
6	25.6	53.9	23.7	21.7
...	...	...	...	...

Scatter plot matrix.



Do you see any particular patterns?

Correlation matrix.

	X1	X2	X3	Y
X1	1.0000000	0.9238425	0.4577772	0.8432654
X2	0.9238425	1.0000000	0.0846675	0.8780896
X3	0.4577772	0.0846675	1.0000000	0.1424440
Y	0.8432654	0.8780896	0.1424440	1.0000000

X_1 and X_2 are correlated, X_1 and X_3 are correlated, X_2 and X_3 are correlated.

Consider the following 4 models.

- Model 1: regression of Y on X_1

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Model 2: regression of Y on X_2

$$Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

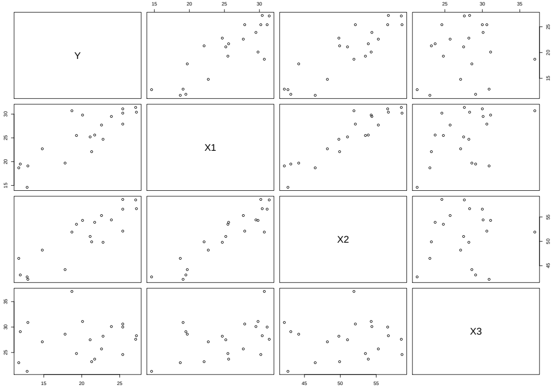
- Model 3: regression of Y on X_1 and X_2

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

Scatter plot matrix.



No obvious nonlinearity.

Correlation matrix.

	X1	X2	X3	Y
X1	1.0000000	0.9238425	0.4577772	0.8432654
X2	0.9238425	1.0000000	0.0846675	0.8780896
X3	0.4577772	0.0846675	1.0000000	0.1424440
Y	0.8432654	0.8780896	0.1424440	1.0000000

X_1 and X_2 are highly correlated, X_1 and X_3 are moderately correlated, X_2 and X_3 are not much correlated.

Consider the following 4 models.

- Model 1: regression of Y on X_1

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Model 2: regression of Y on X_2

$$Y_i = \beta_0 + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Model 3: regression of Y on X_1 and X_2

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

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

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

Boy Fat: Model 1

```
> summary(fit1)

Call:
lm(formula = Y ~ X1, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.4961     3.3192  -0.451    0.658
X1            0.8572     0.1288   6.656 3.02e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.82 on 18 degrees of freedom
Multiple R-squared: 0.7111,     Adjusted R-squared: 0.695
F-statistic:  44.3 on 1 and 18 DF,  p-value: 3.024e-06

> anova(fit1)
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
```

GLT

Boy Fat: Model 2

```
> summary(fit2)

Call:
lm(formula = Y ~ X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -23.6345     5.6574  -4.178 0.000566 ***
X2            0.8565     0.1100   7.786  3.6e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.51 on 18 degrees of freedom
Multiple R-squared: 0.771,      Adjusted R-squared: 0.7583
F-statistic: 60.62 on 1 and 18 DF,  p-value: 3.6e-07

> anova(fit2)
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
```

Boy Fat: Model 3

```
> summary(fit3)

Call:
lm(formula = Y ~ X1 + X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.1742     8.3606  -2.293   0.0348 *
X1            0.2224     0.3034   0.733   0.4737
X2            0.6594     0.2912   2.265   0.0369 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.543 on 17 degrees of freedom
Multiple R-squared: 0.7781,     Adjusted R-squared: 0.7519
F-statistic:  29.8 on 2 and 17 DF,  p-value: 2.774e-06

> anova(fit3)
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
```

GLT

Boy Fat: Model 4

```
> summary(fit4)
Call:
lm(formula = Y ~ X1 + X2 + X3, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X2            -2.857      2.582  -1.106    0.285
X3            -2.186      1.595  -1.370    0.190

Residual standard error: 2.48 on 16 degrees of freedom
Multiple R-squared: 0.8014,     Adjusted R-squared: 0.7641
F-statistic: 21.52 on 3 and 16 DF,  p-value: 7.343e-06

> anova(fit4)
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
```

GLT

Body Fat: ESS

From the R outputs, we can derive a number of extra sums of squares. For example:

- $$SSR(X_2|X_1) = \qquad .$$
- $$SSR(X_1|X_2) = \qquad .$$
- Both extra sums of squares have degrees of freedom , so $MSR(X_2|X_1) =$ and $MSR(X_1|X_2) =$.
- The reduction of SSE by adding to a model with is much more than the reduction of SSE by adding to a model with .

Body Fat: ESS

From the R outputs, we can derive a number of extra sums of squares. For example:

- From Model 1, $SSE(X_1) = 143.12$ and from Model 3, $SSE(X_1, X_2) = 109.95$. So

$$SSR(X_2|X_1) = SSE(X_1) - SSE(X_1, X_2) = 143.12 - 109.95 = 33.17.$$

- From Model 2, $SSE(X_2) = 113.42$, so

$$SSR(X_1|X_2) = SSE(X_2) - SSE(X_1, X_2) = 113.42 - 109.95 = 3.47.$$

- Both extra sums of squares have degrees of freedom 1, so $MSR(X_2|X_1) = 33.17$ and $MSR(X_1|X_2) = 3.47$.
- The reduction of SSE by adding X_2 to a model with X_1 is much more than the reduction of SSE by adding X_1 to a model with X_2.

- $$SSR(X_3|X_1, X_2) = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad .$$

  This extra sum of squares has degrees of freedom ,
  so $MSR(X_3|X_1, X_2) =$.

- $$SSR(X_2, X_3|X_1) = \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad .$$

  This extra sums of squares has degrees of freedom ,
  so $MSR(X_2, X_3|X_1) =$.

Are there other ESS that can be derived from the R outputs?

- From Model 4, $SSE(X_1, X_2, X_3) = 98.40$, so

$$\begin{aligned} SSR(X_3|X_1, X_2) &= SSE(X_1, X_2) - SSE(X_1, X_2, X_3) \\ &= 109.95 - 98.40 = 11.55. \end{aligned}$$

  This extra sum of squares has degrees of freedom 1, so $MSR(X_3|X_1, X_2) = 11.55$.
- Moreover,

  $$SSR(X_2, X_3|X_1) = SSE(X_1) - SSE(X_1, X_2, X_3) = 143.12 - 98.40 = 44.72,$$

  $$SSR(X_1, X_3|X_2) = SSE(X_2) - SSE(X_1, X_2, X_3) = 113.42 - 98.40 = 15.02.$$

  These two extra sums of squares have degrees of freedom 2, so $MSR(X_2, X_3|X_1) = 44.72/2 = 22.36$, $MSR(X_1, X_3|X_2) = 15.02/2 = 7.51$.

Are there other ESS that can be derived from the R outputs?

Decomposition of SSR into ESS

For a model with multiple X variables, the regression sum of squares (SSR) can be expressed as the of several extra sums of squares.

- For example:

$$SSR(X_1, X_2) = \quad .$$

  $SSR(X_1)$ measures the contribution by in the model, whereas $SSR(X_2|X_1)$ measures the contribution when , given that X_1 is already in the model.

- However, such decomposition is usually not unique. For example,

$$SSR(X_1, X_2) = \quad .$$

Decomposition of SSR into ESS

For a model with multiple X variables, the regression sum of squares (SSR) can be expressed as the sum of several extra sums of squares.

- For example:

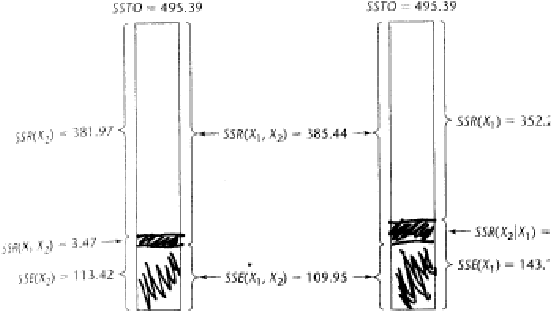
$$SSR(X_1, X_2) = SSR(X_1) + SSR(X_2|X_1).$$

  $SSR(X_1)$ measures the contribution by having X_1 alone in the model, whereas $SSR(X_2|X_1)$ measures the additional contribution when X_2 is added, given that X_1 is already in the model.

- However, such decomposition is usually not unique. For example,

$$SSR(X_1, X_2) = SSR(X_2) + SSR(X_1|X_2).$$

FIGURE 7.1 Schematic Representation of Extra Sums of Squares—Body Fat Example



anova() output

(The next four slides will be discussed on the lab session.)
It provides decomposition of SSR into single d.f. ESS, **in the order of the X variables entering the model.**

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = fat)
> anova(fit4)
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
```

Source of Variation	SS	d.f.	MS
Regression			
Error			
Total			

For example: $SSR(X_2, X_3|X_1) =$

anova() output

It provides decomposition of *SSR* into single d.f. ESS, **in the order of the *X* variables entering the model.**

```
Call:
lm(formula = Y ~ X1 + X2 + X3, data = fat)
> anova(fit4)
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
```

Source of Variation	SS	d.f.	MS
Regression	396.99	3	132.33
X_1	352.27	1	352.27
$X_2 \| X_1$	33.17	1	33.17
$X_3 \| X_1, X_2$	11.55	1	11.55
Error	98.40	16	6.15
Total	495.39	19	

For example: $SSR(X_2, X_3|X_1) = SSR(X_2|X_1) + SSR(X_3|X_1, X_2) = 33.17 + 11.55 = 44.72$.

How to get $SSR(X_2|X_1, X_3)$ from the R output of Model 4? We need to enter the X variables in the following order:

.

```
Call:
lm(formula = Y ~ X1 + X3 + X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X3            -2.186      1.595  -1.370    0.190
X2            -2.857      2.582  -1.106    0.285

> anova(fit4.alt2)
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 ***
X3         1  37.19   37.19  6.0461   0.02571 *
X2         1   7.53    7.53  1.2242   0.28489
Residuals 16  98.40    6.15
```

Then we can get $SSR(X_2|X_1, X_3) =$.

How to get $SSR(X_2|X_1, X_3)$ from the R output of Model 4? We need to enter the X variables in the following order: X_1, X_3, X_2.

```
Call:
lm(formula = Y ~ X1 + X3 + X2, data = fat)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  117.085     99.782   1.173    0.258
X1             4.334      3.016   1.437    0.170
X3            -2.186      1.595  -1.370    0.190
X2            -2.857      2.582  -1.106    0.285

> anova(fit4.alt2)
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 ***
X3         1  37.19   37.19  6.0461   0.02571 *
X2         1   7.53    7.53  1.2242   0.28489
Residuals 16  98.40    6.15
```

Then we can get $SSR(X_2|X_1, X_3) = 7.53$.

General Linear Tests

$\mathcal{I}$ and $\mathcal{J}$ are two non-overlapping index sets.

- **Full model**: contains both $X_{\mathcal{I}}$ and $X_{\mathcal{J}}$.
- Test whether $X_{\mathcal{J}}$ may be dropped out of the full model:

$$H_0 : \beta_j = 0, \text{ for } \textbf{all } j \in \mathcal{J}$$

vs.

$$H_a : \text{some } \beta_j : j \in \mathcal{J} \text{ are nonzero.}$$

- H_0 corresponds to a **reduced model** with only $X_{\mathcal{I}}$.

Basic idea: Compare under the full model with *SSE* under the reduced model by an F ratio:

- Under H_0 (i.e., the model):

$$F^* \sim_{H_0} \qquad .$$

- Reject H_0 at level α if the observed F^* .

Basic idea: Compare *SSE* under the full model with *SSE* under the reduced model by an F ratio:

$$F^* = \frac{\frac{SSE(R)-SSE(F)}{df_R-df_F}}{\frac{SSE(F)}{df_F}} = \frac{MSR(X_{\mathcal{J}}|X_{\mathcal{I}})}{MSE(F)}.$$

- Under H_0 (i.e., the reduced model):

$$F^* \sim_{H_0} F_{df_R-df_F,df_F}.$$

- Reject H_0 at level α if the observed $F^* > F(1-\alpha; df_R - df_F, df_F)$.

Rationale behind the general linear tests.

- If $SSE(F)$ is close to $SSE(R)$, then the additional X variables in the full model to explain the variation in the observations.
  Thus a small $SSE(R) - SSE(F)$ is evidence for

  .
- On the other hand, a large $SSE(R) - SSE(F)$ means that the additional X variables in the full model
  the deviation of the observations around the fitted regression surface, and thus serves as evidence for

  .

Rationale behind the general linear tests.

- If $SSE(F)$ is close to $SSE(R)$, then the additional X variables in the full model do not contribute much to explain the variation in the observations.
  Thus a small $SSE(R) - SSE(F)$ is evidence for H_0, i.e., the reduced model.
- On the other hand, a large $SSE(R) - SSE(F)$ means that the additional X variables in the full model substantially reduce the deviation of the observations around the fitted regression surface, and thus serves as evidence for H_a, i.e., the full model.

F-test for Regression Relation

- Full model with $X_1, \cdots, X_{p-1}$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \cdots n.$$

- Reduced model with no X variable:

$$Y_i = \beta_0 + \epsilon_i, \;\; i = 1, \cdots, n.$$

  So $SSE(R) =$,and $df_R =$.

- $SSE(R) - SSE(F) =$, and $df_R - df_F =$.
- F ratio

$$F^* = \qquad .$$

F-test for Regression Relation

- Full model with $X_1, \cdots, X_{p-1}$:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \cdots n.$$

- Reduced model with no X variable:

$$Y_i = \beta_0 + \epsilon_i, \;\; i = 1, \cdots, n.$$

  So $SSE(R) = SSTO$ and $df_R = n - 1$.

- $SSE(R) - SSE(F) = SSTO - SSE(F) = SSR(F)$, and $df_R - df_F = (n-1) - (n-p) = p - 1 = d.f.(SSR(F))$.
- F ratio

$$F^* = \frac{SSR(F)/(p-1)}{SSE(F)/(n-p)} = \frac{MSR(F)}{MSE(F)}.$$

Test whether a Single $\beta_k = 0$

Body fat: Test for the model with all three predictors whether the midarm circumference (X_3) can be dropped.

- Full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

  $SSE(F) = 98.40$ with d.f. 16.
- Null and alternative hypotheses:

$$H_0 : \qquad\qquad\qquad vs. \quad H_a : \qquad\qquad\qquad .$$

- Reduced model:

  $SSE(R) =$ with d.f. .
- $F^* =$.
- Pvalue= . So we X_3 from the full model.

Test whether a Single $\beta_k = 0$

Body fat: Test for the model with all three predictors whether the midarm circumference (X_3) can be dropped.

- Full model: $SSE(F) = 98.40$ with d.f. 16.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Null and alternative hypotheses:

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

- Reduced model: $SSE(R) = 109.95$ with d.f. 17.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- $F^* = \frac{11.55/1}{98.40/16} = 1.88.$
- Pvalue=$P(F_{1,16} > 1.88) = 0.189$. So we can drop X_3 from the full model.

Equivalence between F-test and T-test

- Test whether X_k can be dropped from a regression model with $p-1$ X variables:
$$H_0 : \beta_k = 0 \;\; vs. \;\; H_a : \beta_k \neq 0.$$
- We can use an F-test: $F^* \underset{H_0}{\sim} F_{1,n-p}$.
- Alternatively, we may use a T-test:
$$T^* = \frac{\hat{\beta}_k}{s\{\hat{\beta}_k\}} \underset{H_0}{\sim} t_{(n-p)},$$
where $\hat{\beta}_k$ is the LS estimator of β_k and $s\{\hat{\beta}_k\}$ is its standard error under the full model.
- It can be show that $F^* = (T^*)^2$ and $F(1-\alpha; 1, n-p) = (t(1-\alpha/2; n-p))^2$. So in this case F-test and T-test are equivalent.

Notes: for one one-sided alternatives, we still need the T-tests.

Test whether Several $\beta_k = 0$

Body fat: Test whether both thigh circumference (X_2) and midarm circumference (X_3) can be dropped from the model with all three predictors.

- Full model: $SSE(F) = 98.40$ with d.f. 16.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Null and alternative hypotheses:

$$H_0 : \qquad\qquad\qquad vs. \quad H_a :$$

- Reduced model: $SSE(R) =$ with d.f. .

- $F^* =$.
- Pvalue= . The result is at $\alpha = 0.05$.

Test whether Several $\beta_k = 0$

Body fat: Test whether both thigh circumference (X_2) and midarm circumference (X_3) can be dropped from the model with all three predictors.

- Full model: $SSE(F) = 98.40$ with d.f. 16.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Null and alternative hypotheses:

$$H_0 : \beta_2 = \beta_3 = 0 \;\; vs. \quad H_a : \text{not both } \beta_2 \text{ and } \beta_3 \text{ equal zero.}$$

- Reduced model: $SSE(R) = 143.12$ with d.f. 18.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- $F^* = \frac{44.72/2}{98.40/16} = 3.635.$
- Pvalue$= P(F_{2,16} > 3.635) = 0.0499$. The result is barely significant at $\alpha = 0.05$.

Test Equality of Several β_ks

- Full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i.$$

- For $q \leq p - 1$:

$$H_0 : \beta_1 = \cdots = \beta_q \text{ vs. } H_a : \beta_1, \cdots, \beta_q \text{ are not all equal.}$$

- Reduced model:

$$Y_i = \beta_0 + \beta_c (X_{i1} + \cdots + X_{iq}) + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i.$$

- β_c denotes the common value of $\beta_1, \cdots, \beta_q$ under H_0, and $X_1 + \cdots + X_q$ is the corresponding (new) X variable. $SSE(R)$ has d.f. $n - (p - q + 1)$.
- $F^* = \frac{(SSE(R) - SSE(F))/(q-1)}{SSE(F)/(n-p)} \underset{H_0}{\sim} F_{q-1,n-p}$.

Body Fat

Test for the model with all three predictors whether the thigh circumference (X_2) and the midarm circumference (X_3) have the same effect.

- Full model: $SSE(F) = 98.40$ with d.f. 16.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Null and alternative hypotheses:

$$H_0 : \qquad\qquad vs. \quad H_a : \qquad\qquad .$$

- Reduced model:

Model 4

Body Fat

Test for the model with all three predictors whether the thigh circumference (X_2) and the midarm circumference (X_3) have the same effect.

- Full model: $SSE(F) = 98.40$ with d.f. 16.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i, \quad i = 1, \cdots, 20.$$

- Null and alternative hypotheses:

$$H_0 : \beta_2 = \beta_3 \;\; vs. \quad H_a : \beta_2 \neq \beta_3.$$

- Reduced model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_c (X_{i2} + X_{i3}) + \epsilon_i, \quad i = 1, \cdots, 20.$$

Model 4

```
> fat.new=data.frame(cbind(fat[,"X1"],fat[,"X2"]+fat[, "X3"], fat[,"Y"]))
> colnames(fat.new)=c("X1",  "X2plusX3","Y")
> fit5=lm(Y~X1+X2plusX3, data=fat.new) ##reduced model
> summary(fit5)
Call:
lm(formula = Y ~ X1 + X2plusX3, data = fat.new)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  52.3706    20.4705   2.558 0.020357 *
X1            2.3732     0.5812   4.083 0.000774 ***
X2plusX3     -1.1706     0.4404  -2.658 0.016573 *
---
Residual standard error: 2.439 on 17 degrees of freedom
Multiple R-squared: 0.7959,     Adjusted R-squared: 0.7719
F-statistic: 33.15 on 2 and 17 DF,  p-value: 1.36e-06

> anova(fit5)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1 352.27  352.27 59.2287 6.16e-07 ***
X2plusX3   1  42.01   42.01  7.0634  0.01657 *
Residuals 17 101.11    5.95
```

- $SSE(R) = 101.11$ with degrees of freedom .
- F ratio:

$$F^* = \quad .$$

- Pvalue$=$.
- The result is and we the null hypothesis that $\beta_2 = \beta_3$. We conclude that the thigh circumference (X_2) and the midarm circumference (X_3) .

- $SSE(R) = 101.11$ with degrees of freedom $17(= 20 - 3)$.
- F ratio:

$$F^* = \frac{(101.11 - 98.40)/(17 - 16)}{98.40/16} = \frac{2.71}{6.15} = 0.44.$$

- Pvalue$= P(F_{(1,16)} > 0.44) = 0.52$.
- The result is not significant and we can not reject the null hypothesis that $\beta_2 = \beta_3$. We conclude that the thigh circumference (X_2) and the midarm circumference (X_3) have the same effect.

Test whether One or Several $\beta_k = \beta_k^{(0)}$

- Full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \cdots n.$$

- For $q \leq p - 1$:

$$H_0 : \beta_1 = \beta_1^{(0)}, \cdots, \beta_q = \beta_q^{(0)} \ vs.\ H_a : \text{not all equalities in } H_0 \text{ hold.}$$

- Reduced model:

- Reduced model has a new response variable
. $SSE(R)$ has d.f. .

- $F^* = \frac{(SSE(R) - SSE(F))/q}{SSE(F)/(n-p)} \underset{H_0}{\sim} F_{q,n-p}$.

Test whether One or Several $\beta_k = \beta_k^{(0)}$

- Full model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i, \quad i = 1, \cdots n.$$

- For $q \leq p - 1$:

$$H_0 : \beta_1 = \beta_1^{(0)}, \cdots, \beta_q = \beta_q^{(0)} \; vs. \; H_a : \text{not all equalities in } H_0 \text{ hold.}$$

- Reduced model: Define $\tilde{Y}_i := Y_i - \sum_{k=1}^{q} \beta_k^{(0)} X_{ik}$

$$\tilde{Y}_i = \beta_0 + \beta_{q+1} X_{i,q+1} + \cdots + \beta_{p-1} X_{i,p-1} + \epsilon_i.$$

- Reduced model has a new response variable $\tilde{Y} = Y - \sum_{k=1}^{q} \beta_k^{(0)} X_k$. $SSE(R)$ has d.f. $n - (p - q)$.
- $F^* = \frac{(SSE(R) - SSE(F))/q}{SSE(F)/(n-p)} \underset{H_0}{\sim} F_{q,n-p}$.