STA302/1001 - Methods of Data Analysis I

(Week 06 lecture A)

Wei (Becky) Lin

June 19-23, 2017

Last Lecture

- Geometric perspective of least squares regression
- F test for regression coefficients.
- Coefficient of Multiple Determination.
- Inferences about Regression Parameters.
- Interval Estimation of β_k, E(Y_h).
- Extra sum of squares

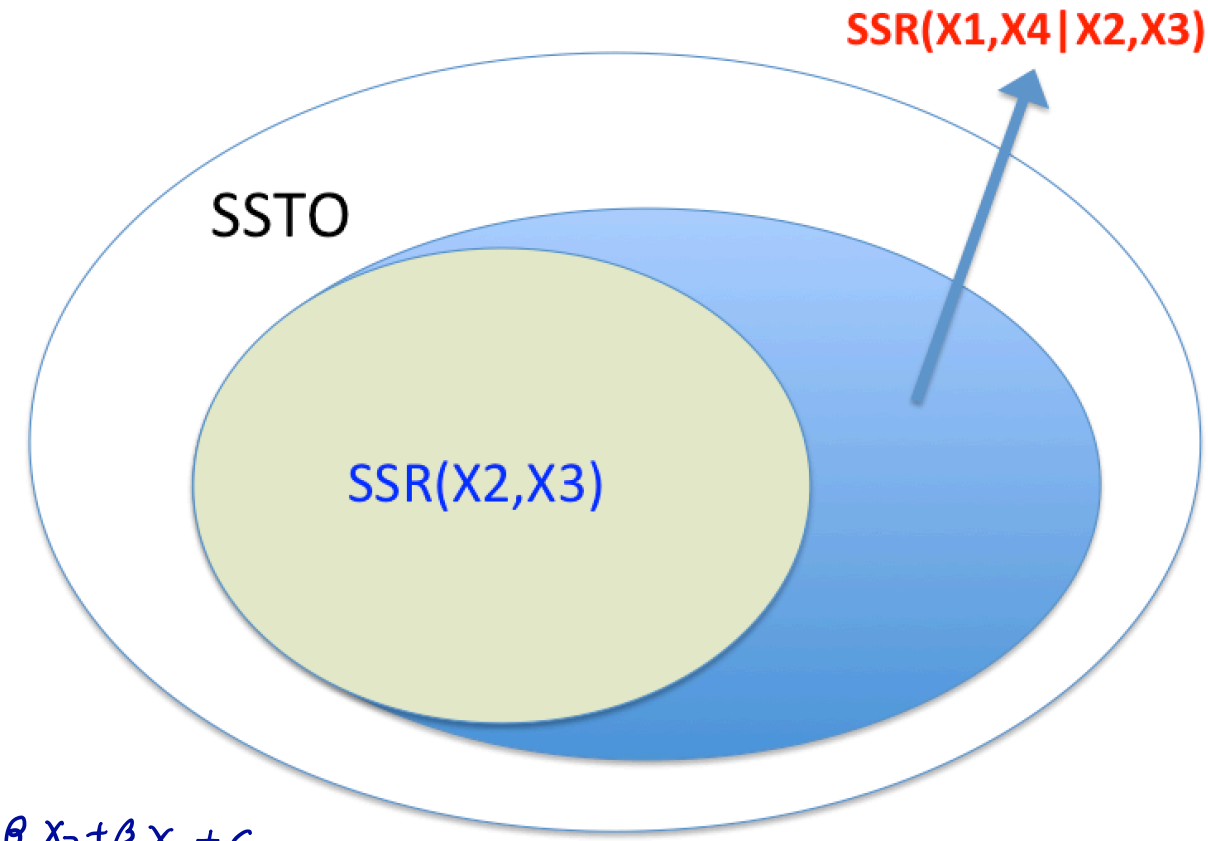
Week 06 Lecture A- Learning objectives & Outcomes

- Review of Extra sum of squares
- Type I and Type III SS
- Use of Extra Sum of Squares in Tests for Regression coefficients
- Coefficient of Partial Determination
- Summary of Tests Concerning Regression Coefficients
- Multicollinearity and Its Effects

Review on Extra Sum of Squares

- Extra Sum of Squares measures the marginal decrease in SSE (equivalently, the marginal increase in SSR) when one or several predictor variables are added to the regression model, given that other variables are already in the model.
- Extra: SSE goes down by the amount of x, SSR goes up by the same amount of x since SSTO=SSR+SSE.
- Examples:
  - $SSR(X_1, X_2, X_3)$ is the total variation explained by X_1X_2, and X_3 in a model
  - $SSR(X_1|X_2)$ is the additional variation explained by X_1 added to a model already containing X_2.
  - $SSR(X_1, X_4|X_2, X_3)$ is the additional variation explained by X_1 and X_4 when they are added to a model already containing X_2 and X_3.
  - Subscripts after bar (|) represent variables already in model.

Review on Extra Sum of Squares

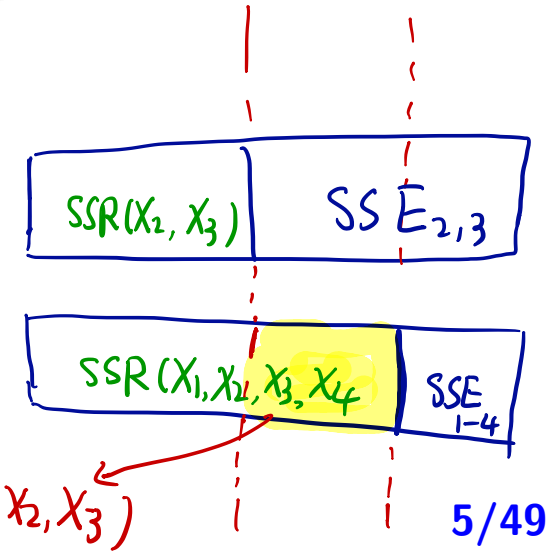


m1: $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

m2: $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_1 X_1 + \beta_4 X_4 + \varepsilon$

SSTO under m1 :

SSTO under m2



$SSR(X_1, X_2, X_3, X_4) - SSR(X_2, X_3)$

$SSE(X_2, X_3) - SSE(X_1, X_2, X_3, X_4)$

$= SSR(X_1, X_4 | X_2, X_3)$

Sequential SS (type I SS)

- SSR (Type I SS) decomposition

Extra sum of squares given previous X's

SSR	df
X_1	1
$X_2 \mid X_1$	1
$X_3 \mid X_1, X_2$	1
(X_1, X_2, X_3)	3

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

□ = □ = same

SSR	df
X_2	1
$X_1 \mid X_2$	1
$X_3 \mid X_1, X_2$	1
(X_1, X_2, X_3)	3

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

- Depends on variable order

Extra Sum of Squares (type I SS)

$$Y = \beta_0 + \beta_1 X_i + \ldots + \beta_{p-1} X_{p-1} + \epsilon$$

- $SSR = \sum_i (\hat{Y}_i - \bar{Y})^2 = SSR(X_1, \ldots, X_{p-1})$
- $SSE = \sum_i (Y_i - \hat{Y}_i)^2 = SSE(X_1, \ldots, X_{p-1})$
- Extra Sum of Squares
  - Break down the SSR to contributions from different X's sequentially
  - $SSR(X_1), SSR(X_2|X_1), SSR(X_3|X_1, X_2), \ldots$
    - $SSR(X_2|X_1) = SSR(X_1, X_2) - SSR(X_1) = SSE(X_1) - SSE(X_1, X_2)$
  - $SSR = SSR(X_1) + SSR(X_2|X_1) + \ldots SSR(X_{p-1}|X_1, \ldots, X_{p-2})$
  - Degrees of freedom of extra SSR equal to number of extra variables
    - e.g. $df_{SSR(X_3|X_1)} = 1, df_{SSR(X_2,X_3|X_1)} = 2$
  - Contributions are not additive

$$SSR(X_1, X_2, X_3) \neq SSR(X_1) + SSR(X_2) + SSR(X_3)$$

Extended ANOVA Table

R output

Type I ss

Source of Variance	SS	df	MS
X_1	$\text{SSR}(X_1)$	1	$\text{MSR}(X_1)$
$X_2 \mid X_1$	$\text{SSR}(X_2 \mid X_1)$	1	$\text{MSR}(X_2 \mid X_1)$
$X_3 \mid X_1, X_2$	$\text{SSR}(X_3 \mid X_1, X_2)$	1	$\text{MSR}(X_3 \mid X_1, X_2)$
Error	$\text{SSE}(X_1, X_2, X_3)$	n-4	$\text{MSR}(X_1, X_2, X_3)$
Total	SSTO	n-1	

default anova output in R using anova( )

F test in anova output (type I SS)

```
SSR2=sum(anova(m2)[-3,2]) # SSR(X1,X2)=385.44, SSE(X1,X2)=109.95
m3 =lm(Y~X1+X2+X3,data=body); anova(m3)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value      Pr(>F)
## X1         1 352.27  352.27 57.2768 0.000001131 ***
## 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
```

```
SSR3 =sum(anova(m3)[-4,2]) # SSR(X1,X2,X3)=396.98, SSE(X1,X2,X3)=98.40
```

- Type I SS: variables added in order. Sum of sequential SSR gives SSR.
- F-tests are testing each variable given previous variables already in model.

Type III/II SS

- SSR (Type III SS) decomposition: refers to variable added last.
- These do NOT add to the SSR.
- F-tests are testing variable given that all of the other variables already in the model.

	SSR	df
X_1:	$X_1 \mid X_2, X_3$	1
X_2:	$X_2 \mid X_1, X_3$	1
X_3:	$X_3 \mid X_1, X_2$	1
	(X_1, X_2, X_3)	3

$X_k \mid X_{-k}$: given all the rest X's

sum $SSR(X_k \mid X_{-k}) \neq SSR$

- Does not depend on variable order
- Type II SS are pretty much the same as Type III, except they ignore interaction terms.

Type I vs Type III

- Estimates using Type I SS tell us how much of the variation in Y can be explained by X_1, how much of the residual variability ($SSE(X_1)$) can be explained by X_2, how much of the remaining residual ($SSE(X_1, X_2)$) can be explained by X_3 and so on, in order.

- Estimates using Type III SS tell us how much of the residual variability in Y can be accounted for by X_1 after having accounted for everything else, and how much of the residual variability in Y can be accounted for X_2 after having accounted for everything else as well, and so on.

7.2 Use of Extra Sums of Squares in Tests for Regression Coefficients

Partial F test: Test whether several $\beta_k = 0$

- Consider a regression model, we call Full model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_{q+p} X_{q+p} + \epsilon$$

- We want to test the null hypothesis that some of the β_k are zero

$$H_0 : \beta_{q+1} = \ldots = \beta_{q+p} = 0$$

- The alternative hypothesis is

$$H_A : \text{not all } \beta_{q+1}, \ldots, \beta_{q+p} \text{ equal zero}$$

- The general linear test approach:

$$F^* = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F}$$

  - Decision: reject H_0, in favor of H_a at α significance level if

$$F^* \geq F_{1-\alpha; df_R - df_F, df_F}$$

Partial F test: Test whether several $\beta_k = 0$ (contd.)

- Full model:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \beta_{q+1} X_{q+1} + \ldots + \beta_{q+p} X_{q+p} + \epsilon$$

- Reduced model (under H_0):

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_q X_q + \epsilon$$

- From Full model, we get:

$$SSR(X_1, \ldots, X_q, \ldots, X_{q+p}), \quad SSE(X_1, \ldots, X_q, \ldots, X_{q+p})$$

- From Reduced model, we get:

$$SSR(X_1, \ldots, X_q), \quad SSE(X_1, \ldots, X_q)$$

Partial F test: Test whether several $\beta_k = 0$ (contd.)

- The extra sum of squares is obtained as

$$SSR(X_{q+1}, \ldots, X_{p+q} | X_1, \ldots, X_q) = SSE_R - SSE_F$$
$$= SSE(X_1, \ldots, X_q) - SSE(X_1, \ldots, X_q, \ldots, X_{q+p})$$

- Alternatively

$$SSR(X_{q+1}, \ldots, X_{p+q} | X_1, \ldots, X_q) = SSR_F - SSR_R$$
$$= SSR(X_1, \ldots, X_q, \ldots, X_{q+p}) - SSR(X_1, \ldots, X_q)$$

- Hence, the test statistic is

$df_R = n - (q+1)$
$df_F = n - (q+p+1)$
$p = df_R - df_F$

$$F^* = \frac{(SSE_R - SSE_F)/p}{SSE(X_1, \ldots, X_{p+q})/(n-p-q-1)}$$
$$= \frac{SSR(X_{q+1}, \ldots, X_{p+q} | X_1, \ldots, X_q)/p}{SSE(X_1, \ldots, X_{p+q})/(n-p-q-1)}$$

- Decision: Reject H_0, conclude H_a at α significance level if

$$F^* \geq F_{1-\alpha; p, n-p-q-1}$$

Body fat Example:Testing a single $\beta_3 = 0$

$$\textit{Full model} \quad Y_i = \beta_i + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- Test:

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

- Reduced model under H_0:

$$\textit{Reduced model :} \quad Y_i = \beta_i + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

- The test whether or not $\beta_3 = 0$ is a marginal test, given X_1, X_2 are already in the the model.

Body fat Example:Testing a single $\beta_3 = 0$ (contd.)

- $SSE_F = SSE(X_1, X_2, X_3), df = n - 4$
- $SSE_R = SSE(X_1, X_2), df = n - 3$
- The general linear test statistics

$$
\begin{aligned}
F^* &= \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} \\
&= \frac{[SSE(X_1, X_2) - SSE(X_1, X_2, X_3)]/[(n-3) - (n-4)]}{SSE(X_1, X_2, X_3)/(n-4)} \\
&= \frac{SSR(X_3|X_1, X_2)/1}{SSE(X_1, X_2, X_3)/(n-4)} \\
&= \frac{MSR(X_3|X_1, X_2)}{MSE_F}
\end{aligned}
$$

Body fat Example:Testing a single $\beta_3 = 0$ (contd.)

```
SSR2=sum(anova(m2)[-3,2]) # SSR(X1,X2)=385.44, SSE(X1,X2)=109.95
m3 =lm(Y~X1+X2+X3,data=body); anova(m3)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value      Pr(>F)
## X1         1 352.27  352.27 57.2768 0.000001131 ***
## 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
```

MSE_F $MSR(X_3|X_1, X_2)$

```
SSR3 =sum(anova(m3)[-4,2]) # SSR(X1,X2,X3)=396.98, SSE(X1,X2,X3)=98.40
```

- Test statistics:
  $F^* = MSR(X_3|X_1, X_2)/MSE_F = 11.54/6.15 = 1.876423$
- Decision: $F^* = 1.876 \leq 4.494 = F_{1-0.05,1,16}$, failed to reject H_0.

Body fat Example:Testing a single $\beta_k = 0$ (contd.)

```
body=read.table("/Users/Wei/TA/Teaching/0-STA302-2016F/Week10-Nov14/bodyfat.txt",header=T)
n <- dim(body)[1]
fmod <- lm(Y~. ,data=body)
anova(fmod)
```

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

type I SS

```
## 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
```

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

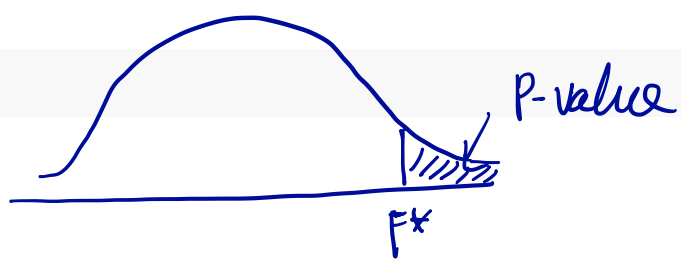
Method 1:

$$F^* = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F}$$

```
rmod <- lm(Y~X1+X2,data=body)
SSEf = deviance(fmod)
SSEr = deviance(rmod)
Ft <- ((SSEr-SSEf)/1)/(SSEf/(n-4))
Ft
```

```
## [1] 1.877289
```

```
pf(Ft,1,n-4,lower.tail=F)
```




```
## [1] 0.1895628
```

Body fat Example:Testing a single $\beta_3 = 0$ (contd.)

```
body=read.table("/Users/Wei/TA/Teaching/0-STA302-2016F/Week10-Nov14/bodyfat.txt",header=T)
n <- dim(body)[1]
fmod <- lm(Y~. ,data=body)
rmod <- lm(Y~X1+X2,data=body)
anova(rmod,fmod)
```

→ Full model

→ Reduced model

```
## Analysis of Variance Table
##
## Model 1: Y ~ X1 + X2
## Model 2: Y ~ X1 + X2 + X3
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     17 109.951
## 2     16  98.405  1    11.546 1.8773 0.1896
```

↑ F^*

↑ $P(F_{1,16} > F^*) = 0.1896$

Body fat Example:Testing $\beta_1 = \beta_3 = 0$

$$H_0 : \beta_1 = \beta_3 = 0 \quad H_a : \textit{not both } \beta_1, \beta_3 \textit{ equal zero.}$$

- Full model:

$$Y_i = \beta_i + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

- Reduced model (under H_0):

$$Y_i = \beta_i + \beta_2 X_2 + \epsilon$$

- Test statistics

$$\begin{aligned} F^* &= \frac{[SSE(X_2) - SSE(X_1, X_2, X_3)]/[(n-2) - (n-4)]}{SSE(X_1, X_2, X_3)/(n-4)} \\ &= \frac{SSR(X_1, X_3|X_2)/2}{SSE_F/(n-4)} \\ &= \frac{MSR(X_1, X_3|X_2)}{MSE_F} \end{aligned}$$

Body fat Example:Testing $\beta_1 = \beta_3 = 0$

```
body=read.table("/Users/Wei/TA/Teaching/0-STA302-2016F/Week10-Nov14/bodyfat.txt",header=T)
n <- dim(body)[1]; fmod <- lm(Y~X2+X1+X3,data=body)
anova(fmod)
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value    Pr(>F)
## X2         1 381.97  381.97 62.1052 6.735e-07 ***
## X1         1   3.47    3.47  0.5647    0.4633
## X3         1  11.55   11.55  1.8773    0.1896
## Residuals 16  98.40    6.15
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_1, X_3 | X_2)$
$= SSR(X_1|X_2) + SSR(X_3|X_1, X_2)$
$= 3.47 + 11.55$
$= 15.02$

```
sum(anova(fmod)[2:3,2])/2/anova(fmod)[4,3]
```

$= MSE_F$

$SSR(X_1, X_3|X_2)/2$

```
## [1] 1.22098
```

```
rmod <- lm(Y~X2,data=body)
anova(rmod,fmod)
```

$$F^* = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{MSE_F}$$

```
## Analysis of Variance Table
##
## Model 1: Y ~ X2
## Model 2: Y ~ X2 + X1 + X3
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1     18 113.424
## 2     16  98.405  2    15.019 1.221  0.321
```

Comments

- Testing whether a single β_k equals zero:
  - the t^* test statistic
  - the F^* general linear test statistic
  - $F^* = (t^*)^2$ when X_k is the last predictor in the full model using Type I SS.
  - $F^* = (t^*)^2$ for $\forall k$ when use Type III SS.

refer slides 25 & 19

refer slides 25/26

- Testing whether several β_k equal zero:
  - the F^* general linear test statistic (partial F test)
- General linear test statistic can be expressed in term of the the coefficients of multiple determination R^2

$$F^* = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} = \frac{(R_F^2 - R_R^2)/(df_R - df_F)}{(1 - R_F^2)/df_F}$$

  - The latter formula using R^2 is not appropriate when the full and reduced models do not contain β_0

Show

$$F^* = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} = \frac{(R_F^2 - R_R^2)/(df_R - df_F)}{(1 - R_F^2)/df_F}$$

- For given Y, SSTO are the same for full model and reduced model.

- $$F^* = \frac{\frac{SSE_R - SSE_F}{df_R - df_F} / SSTO}{\frac{SSE_F}{df_F} / SSTO}$$

$$\frac{SSE_R}{SSTO} = 1 - R_R^2$$

$$\frac{SSE_F}{SSTO} = 1 - R_F^2$$

$$= \frac{R_F^2 - R_R^2}{df_R - df_F} \div \frac{1 - R_F^2}{df_F}$$

Body Fat example: $F^* = (t^*)^2$ using Type III SS

```
body=read.table("/Users/Wei/TA/Teaching/0-STA302-2016F/Week10-Nov14/bodyfat.txt",header=T)
fmod <- lm(Y~X1+X2+X3,data=body)
summary(fmod)
```

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

```
(summary(fmod)$coef[,"t value"])^2
```

$\leftarrow (t^*)^2$

```
## (Intercept)          X1          X2          X3
##    1.376868    2.065734    1.224212    1.877289
```

$\rightarrow$ same as F-value in next slide

Body Fat example: $F^* = (t^*)^2$ using Type III SS

```
body=read.table("/Users/Wei/TA/Teaching/0-STA302-2016F/Week10-Nov14/bodyfat.txt",header=T)
fmod <- lm(Y~X1+X2+X3,data=body)
library(car)
Anova(fmod,type=3)
```

```
## Anova Table (Type III tests)
##
## Response: Y
##             Sum Sq Df F value Pr(>F)
## (Intercept)  8.468  1  1.3769 0.2578
## X1          12.705  1  2.0657 0.1699
## X2           7.529  1  1.2242 0.2849
## X3          11.546  1  1.8773 0.1896
## Residuals   98.405 16
```

type III SS

```
sqrt(Anova(fmod,type=3)[1:4,3])
```

```
## [1] 1.173400 1.437266 1.106441 1.370142
```

7.3 Summary of Tests concerning Regression coefficients

Summary

- Test whether all $\beta_k = 0$
  - overall F test:

$$F^* = \frac{MSR}{MSE} \sim F_{p-1,n-p}$$

- Test whether a single $\beta_k = 0$
  - Partial F test:

$$F^* = \frac{MSR(X_k|X_1,\ldots,X_{k-1},X_{k+1},\ldots,X_{p-1})}{MSE} \sim F_{1,n-p}$$

  - $F^* = (t^*)^2 = \frac{b_k}{s\{b_k\}}$ true for last predictor using Type I SS, and true for any k using type III SS.

Summary (contd.)

- Test whether some $\beta_k = 0$

  $H_0 : \beta_q = \beta_{q+1} = \ldots = \beta_{p-1} = 0$, H_a : not all β_k in H_0 equal zero.

  - partial F test:

$$F^* = \frac{(SSE_R - SSE_F)/(df_R - df_F)}{SSE_F/df_F} = \frac{(R_F^2 - R_R^2)/(df_R - df_F)}{(1 - R_F^2)/df_F}$$
$$= \frac{MSR(X_q, \ldots, X_{p-1}|X_1, \ldots, X_{p-2})}{MSE_F} \sim F_{p-q,n-p}$$

Summary (contd.)

- Other test using general linear test

$$Full: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \quad \leftarrow df_F = n-4$$

- $H_0 : \beta_1 = 2\beta_2, H_a : \beta_1 \neq 2\beta_2$
- Reduced: $Y = \beta_0 + \beta_c(2X_1 + X_2) + \beta_3 X_3 + \epsilon \quad \leftarrow df_R = n-3$
- The general F^* test statistics $\sim F_{1,n-4}$

$$df_R - df_F = (n-3) - (n-4) = 1$$

- $H_0 : \beta_1 = 3; \beta_3 = 5, H_a$: not both equalities in H_0 hold
- Reduced: $Y - 3X_1 - 5X_3 = \beta_0 + \beta_2 X_2 + \epsilon \quad \leftarrow df_R = n-2$
- The general F^* test statistics $\sim F_{2,n-4}$

$$df_R - df_F = (n-2) - (n-4) = 2$$

7.4 Coefficient of Partial Determination

Coefficient of Partial Determination

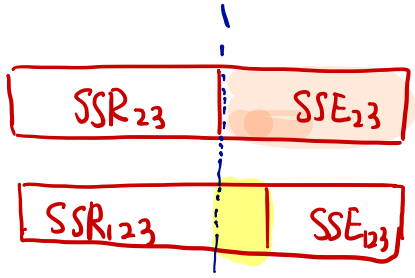
- Coefficient of determination

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

  - the percentage of the total variation in Y that has been explained by the model.

- Partial Determination: the amount of remaining variation explained by a variable given other variables already in the model, this is called partial determination.
- Coefficient of Partial Determination:

$$R^2_{Y\,1|23} = \frac{SSE(X_2, X_3) - SSE(X_1, X_2, X_3)}{SSE(X_2, X_3)} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)}$$



$$R^2_{Y\,1|23} = \frac{SSR(X_1|X_2,X_3)}{SSE(X_2,X_3)} = \frac{\quad}{\quad} = \frac{SSE_R - SSE_F}{SSE_R}$$

Coefficient of Partial Determination (contd.)

- Full model: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$
- Reduced model: $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \epsilon$

$$R^2_{Y\ 1|23} = \frac{SSE(X_2, X_3) - SSE(X_1, X_2, X_3)}{SSE(X_2, X_3)} = \frac{SSR(X_1|X_2, X_3)}{SSE(X_2, X_3)}$$

SSE_R SSE_F SSE_R

- Measure relative reduction in Y variance after introducing X_1 to model with X_2, X_3.
- Takes values in [0,1]
- $R^2_{Y\ 1|23} = R^2$ of regressing residuals of reduced model to residuals of $E(X_1) = \beta_0 + \beta_1 X_2 + \beta_2 X_3$.

- m1 = lm(Y ~ X2 + X3)
- m2 = lm(X1 ~ X2 + X3)
- lm(m1$resid ~ m2$resid) → $R^2 = R^2_{Y\ 1|23}$

Coefficient of Partial Determination (contd.)

- $R^2_{Y\ 1|23} = R^2$ of regressing residuals of reduced model to residuals of $E(X_1) = \beta_0 + \beta_1 X_2 + \beta_2 X_3$.
  - Regress Y on X_2, X_3 to get $\hat{Y}_i(X_2, X_3)$ and $e_i(Y|X_2, X_3)$
  - Regress X_1 on X_2, X_3 to get $\hat{X}_i(X_2, X_3)$ and $e_i(X_1|X_2, X_3)$
  - R^2 between $e_i(Y|X_2, X_3)$ and $e_i(X_1|X_2, X_3)$ will be the same as $R^2_{Y\ 1|23}$.
  - added variable plots or partial regression plot: the strength of the relationship between Y and X_1 adjusted for X_2, X_3.

$$e_i(Y|X_2, X_3) \text{ vs } e_i(X_1|X_2, X_3)$$

- More generally

$$R^2_{Y\ p,\ldots,m|1,2,\ldots,p-1} = \frac{SSR(X_p, \ldots, X_m|X_1, \ldots, X_p, \ldots X_m)}{SSE(X_1, \ldots, X_p, \ldots, X_m)} = \frac{SSE_R - SSE_F}{SSE_R}$$

- Coefficient of Partial Correlation

$$r_{Y\ k|1,\ldots,p-1} = sign(b_k)\sqrt{R^2_{Y\ k|1,\ldots,p-1}}$$

7.6 Multicollinearity and its effect

What is multicollinearity

- Multicollinearity, is also called collinearity or intercorrelation: the predictor variables are correlated among themselves.
- Uncorrelated predictor variables: the marginal reduction in the SSE when the other predictor variables are in the model is exactly the same when the predictor variable is in the model.
    - eg. X_1, X_2 are uncorrelated, then $SSR(X_1|X_2) = SSR(X_1), SSR(X_2|X_1) = SSR(X_2)$

Uncorrelated predictors

```
X1 = c(rep(4,4),rep(6,4))  # crew size
X2=c(2,2,3,3,2,2,3,3)       # bonus pay
Y=c(42,39,48,51,49,53,61,60)  # crew productivity
cor(X1,X2)
```

```
## [1] 0
```

X_1, X_2 : uncorrelated

```
anova(lm(Y~X1+X2))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X1         1 231.125 231.125  65.567 0.0004657 ***
## X2         1 171.125 171.125  48.546 0.0009366 ***
## Residuals  5  17.625   3.525
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_2|X_1) = 171.125$

Uncorrelated predictors

```
X1 = c(rep(4,4),rep(6,4))  # crew size
X2=c(2,2,3,3,2,2,3,3)       # bonus pay
Y=c(42,39,48,51,49,53,61,60)  # crew productivity
cor(X1,X2)
```

```
## [1] 0
```

```
anova(lm(Y~X1))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X1         1 231.12 231.125   7.347 0.03508 *
## Residuals  6 188.75  31.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(lm(Y~X2))
```

```
## Analysis of Variance Table
##
## Response: Y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## X2         1 171.12 171.125  4.1276 0.08846 .
## Residuals  6 248.75  41.458
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$SSR(X_2) = SSR(X_2|X_1)$ ← slide 37

Predictors are perfect correlated

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

Case i	X_{i1}	X_{i2}	Y_i	Fitted Values for Regression Function (7.58)	Fitted Values for Regression Function (7.59)
1	2	6	23	23	23
2	8	9	83	83	83
3	6	8	63	63	63
4	10	10	103	103	103

Response Functions:

$\hat{Y} = -87 + X_1 + 18X_2$ (7.58)

$\hat{Y} = -7 + 9X_1 + 2X_2$ (7.59)

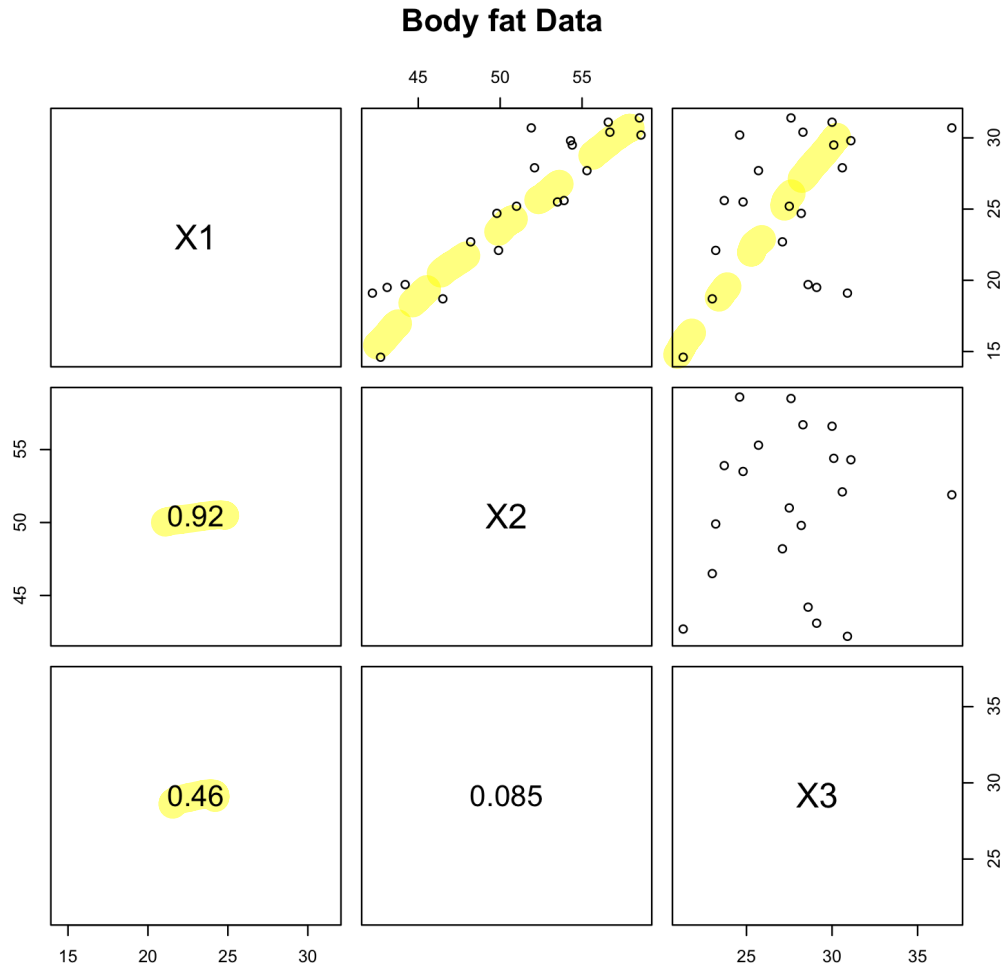
$$X_2 = 5 + 0.5X_1$$

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2(5 + 0.5X_1) = (\beta_0 + 5\beta_2) + (\beta_1 + 0.5\beta_2)X_1$$

Predictors are perfect correlated

- When two predictor variables are perfectly correlated, many response functions will lead to the same fitted values for the observations.
- The perfect relation between X_1 and X_2 did not inhabit the ability to obtain a good fit to the data.
- Since many different response functions provided the same good fit, we can not interpret any one set of regression coefficients as reflecting the effects of the different predictor variables.

Collinearity and its effects: body fat data



Collinearity and its effects: body fat data

- Collinearity effect on regression coefficients

	Variables in model	b_1 (coef of X_1)	b_2 (coef. of X_2)
M1:	X_1	0.8572(***)	-
M2:	X_2	-	0.8565(***)
M3:	X_1, X_2	0.2224	0.6594 (*)
M4:	X_1, X_2, X_3	4.334	-2.857

p.val < 0.01 strong evidence

moderate evidence 0.01 < p.val < 0.05

M1: $Y \sim X_1$

M2: $Y \sim X_2$

M3: $Y \sim X_1 + X_2$

M4: $Y \sim X_1 + X_2 + X_3$

Collinearity and its effects: body fat data

- Collinearity effect on $s(b_k)$

Variables in model	$s(b_1)$	$s(b_2)$
X_1	0.1288	-
X_2	-	0.1100
X_1, X_2	0.3034	0.2912
X_1, X_2, X_3	3.016	2.582

The high degree of multicollinearity among the predictor variables is responsible for the inflated variability of the estimated regression coefficients.

Collinearity and its effects: body fat data

- Collinearity effect on fitted values and predictions

Variables in model	MSE
X_1	7.95
X_1, X_2	6.47
X_1, X_2, X_3	6.15

- Estimated means and predicted values are not affected.

Collinearity and its effects: body fat data

- Collinearity effect on simultaneous tests of β_k
  - it is possible that when individual t tests are performed, neither β_1 or β_2 is significant.
  - However, when the F test is performed for both, the results may still be significant.
- Need for more powerful diagnostics for multicollinearity.

Summary of Multicollinearity

- When X's are orthogonal, $X_k' X_j = 0$, so $X^T X$ is a diagonal matrix
  - parameter estimates are independent, $s^2(b) = MSE(X^T X)^{-1}$ is also diagonal.
  - marginal contribution of each X is additive

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

  - Type I and Type III SS are equivalent.

- When X's are collinear, $X_k = \sum_{j \neq k} a_j X_j$
  - Different set of parameters identical mean response function
  - marginal contribution of each X depends on which of the other variables are already in model
  - The $X^T X$ matrix is not invertible.

Summary of Multicollinearity

- Effects of multicollinearity
  - b's are highly correlated. $cor(b_k, b_j) \approx 1$
  - b's have high variance: $s(b_k)$ is high.
    - individual estimates appear insignificant
  - signs of b's contrary to intuition.
  - problems with parameter interpretation
    - b_k is the rate change in E(Y) per unit change in X_k, keeping other X's fixed, but $X's$ change with X_k.
  - inference for E(Y) and predictions remains valid.
  - Type I and Type III SS are different, except for the last Type I SS.

VIF: Diagnostic for multicollinearity

- Correlation transformation
  - standardized X_k

$$X_k^* = \frac{X_k - \bar{X}_k}{s_k \sqrt{n-1}}$$

- Correlation matrix of $\mathbf{X}$

$$X^{*'} X^* = r_{XX}$$

- The inverse of the correlation matrix

$$r_{XX}^1 = (X^{*'} X^*)^{-1}$$

VIF: Diagnostic for multicollinearity

- VIF: Variance Inflation Factors

$$VIF_k = [r_{XX}^{-1}]_{k,k}, k = 1, \ldots, p-1$$

  - k^{th} diagonal element of r_{XX}^{-1}
- Alternative definition

$$VIF_k = \frac{1}{1 - R_k^2}$$

  - R_k^2 is R^2 for $E(X_k) = \sum_{j \neq k} \beta_j X_j$
- VIF_k measures how much bigger is $S^2(b_k)$ as compared to a model with independent X's.

Use of VIF

- If X's are linearly independent, then $VIF_k \approx 1$
- If X's are collinear, then $VIF >> 1$
- Rule of thumb: if $\max_{k=1,\ldots,p-1}(VIF_k) > 10$, then multicollinearity.

Practice problems after lectures

- keep trying all problems that we have covered today in Ch7:
  - 7.2, 7.3, 7.7, 7.8, 7.12, 7.15, 7.20, 7.22, 7.23, 7.27, 7.31.
- Upcoming topic:
  - model selection.
  - Final review