Coefficients of Partial Determination

- **Definitions.**

  1. Coefficient of Multiple Determination: $R^2 = SSR/SSTO = 1 - SSE/SSTO$.

  2. Coefficient of Partial Determination with 2 Predictors:
  $R^2_{1|2} = SSR(X_1|X_2)/SSE(X_2)$.
  $R^2_{2|1} = SSR(X_2|X_1)/SSE(X_1)$.

  3. Coefficient of Partial Determination with 3 Predictors:
  $R^2_{1|23} = SSR(X_1|X_2, X_3)/SSE(X_2, X_3)$.
  $R^2_{2|13} = SSR(X_2|X_1, X_3)/SSE(X_1, X_3)$.
  $R^2_{3|12} = SSR(X_3|X_1, X_2)/SSE(X_1, X_2)$.

- **Body Fat Example.** The values stored in 'BodyFat.csv' file contains the data for a study of the relation of amount of body fat (Y) to several possible predictor variables, based on a sample of 20 healthy females 25-34 years old. The possible predictor variables are *triceps skinfold thickness* (X_1), *thigh circumference* (X_2), and *midarm circumference* (X_3). The amount of body fat for each of the 20 persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictor variables could provide reliable estimates of the amount of body fat since the measurements needed for the predictor variables are easy to obtain.

```
data.body=read.csv("BodyFat.csv",header=T)
attach(data.body)
y=fat
x1=triceps
x2=thigh
x3=midarm

anova(results.12)
          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

results.1=lm(y~x1)
anova(results.1)
          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

# R^2(2|1)=ssr(x2|x1)/sse(x1)=33.17/143.12=0.232
# Hence, the SSE(x1) is reduced by 23.2 percent.

anova(results.123)
          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

# R^2(3|12)=ssr(x3|x1,x2)/sse(x1,x2)=11.55/109.95=0.105
```

- **Standardized Regression Model**
  - We use this method when det(X'X) is close to zero OR
  - When explanatory variables differ substantially in order of magnitude
- **Dwaine Studios Example.** Dwaine Studios, Inc, operates portrait studios in 21 cities of medium size. These studios specialize in portraits of children. The company is considering an expansion into other cities of medium size and wishes to investigate whether sales (Y) in a community can be predicted from the number of persons aged 16 or younger in the community (X_1) and the per capita disposable personal income in the community (X_2). The data are stored in 'DwaineStudios.csv' file.

```
data=read.csv("DwaineStudios.csv",header=T)
attach(data)
x1=young
x2=income
y=sales
n=length(sales)

x1.star=((x1-mean(x1))/sd(x1))/sqrt(n-1)
x2.star=((x2-mean(x2))/sd(x2))/sqrt(n-1)
y.star=((y-mean(y))/sd(y))/sqrt(n-1)

results.star=lm(y.star~0+x1.star+x2.star)
#coef(results.star)
#  x1.star   x2.star
#0.7483670 0.2511039

b1=(sd(y)/sd(x1))*0.7484
b2=(sd(y)/sd(x2))*0.2511
b0=mean(y)-b1*mean(x1)-b2*mean(x2)

#check
lm(y~x1+x2)
```

- **Multicollinearity.** When the predictor variables are correlated among themselves, *intercorrelation* or *multicollinearity* among them is said to exist.
- **Uncorrelated Predictor Variables.**

```
# Uncorrelated predictors
x1=c(4,4,4,4,6,6,6,6)
x2=c(2,2,3,3,2,2,3,3)
cor(x1,x2)
y=c(42,39,48,51,49,53,61,60)
anova(lm(y~x1))
anova(lm(y~x2))
anova(lm(y~x1+x2))   # Note that ssr(x2)=ssr(x2|x1)
```

- **Perfectly Correlated Predictor Variables.**

```
# Perfectly correlated predictors
x3=c(2,8,6,10)
x4=c(6,9,8,10)
cor(x3,x4)
coefficients(lm(x4~x3))
y2=c(23,83,63,103)

# y2_hat1=-87+x3+18*x4
# y2_hat2=-7+9*x3+2*x4
# y2_hat3=-17+8*x3+4*x4 # Note that all 3 models give perfect fit.
```

- **Body Fat Example.** The values stored in 'BodyFat.csv' file contains the data for a study of the relation of amount of body fat (Y) to several possible predictor variables, based on a sample of 20 healthy females 25-34 years old. The possible predictor variables are *triceps skinfold thickness* (X_1), *thigh circumference* (X_2), and *midarm circumference* (X_3). The amount of body fat for each of the 20 persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictor variables could provide reliable estimates of the amount of body fat since the measurements needed for the predictor variables are easy to obtain.

```
data.body=read.csv("BodyFat.csv",header=T)
attach(data.body)
y=fat
x1=triceps
x2=thigh
x3=midarm

coefficients(lm(y~x1))
coefficients(lm(y~x2))
coefficients(lm(y~x1+x2)) # Note how the beta estimates drastically changes as you
coefficients(lm(y~x1+x2+x3)) # include highly correlated predictors

# SSR(x1|x2) and R^2(1|2)
anova(lm(y~x2+x1)) # Note how small the marginal contribution of x1
                        # when x2 is already in the model.

# SSR(x1|x2) can sometimes be bigger than SSR(x1)
x1=c(5,10,5,10)
x2=c(25,30,5,10)
y=c(20,20,0,1)
temp=data.frame(y=y,x1=x1,x2=x2)
cor(temp)
anova(lm(y~x2))
anova(lm(y~x1+x2)) # x1 is a suppressor variable

# Predictions are still good even with multicollinearity of predictors
x1=triceps
x2=thigh
x3=midarm
y=fat
results=lm(y~x1)
predict(results,new=data.frame(x1=25),interval="confidence",se.fit=T)

results2=lm(y~x1+x2)
predict(results2,new=data.frame(x1=25,x2=50),interval="confidence",se.fit=T)

results3=lm(y~x1+x2+x3)
predict(results3,new=data.frame(x1=25,x2=50,x3=29),interval="confidence",se.fit=T)
```

- **Remedial Measures.**

  1. Restrict the use of the fitted regression model to inferences for values of the predictor variables that follow the same pattern of multicollinearity.
  2. One or several predictor variables may be dropped from the model.
  3. Sometimes it is possible to add some cases that break the pattern of multicollinearity.
  4. In some economic studies, it is possible to estimate the regression coefficients for different predictor variables from different sets of data.
  5. The methodology of *Principal Components* can be used to obtained uncorrelated predictors.
  6. Some transformations of the variables might remove or lessen the pattern of multicollinearity.