ORIGIN ≡ 0

Linear Modeling & "Extra" Sums of Squares

Multiple Regression forms the basis for the extensive field called "Linear Modeling". The independent variables in any particular "linear model" can be many things, including squared X^2 or higher-order terms X^3 or X^4 etc. as in "Polynomial Regression", crossed-factors $X_1 \cdot X_2$ for interactions as seen in ANOVA models, or many kinds of transformed variables for curvilinear regression fits such as ln(X), e^X, or the like. The only requirement is that there is only a single Dependent variable Y, and all Independent variables X_i inter into regression as "linear factors" meaning that they are only added together: $Y=X_1+X_2+X_3$... - even though any Independent variable may be transformed or crossed composites of other variables.

In constructing one or more linear models, the researcher is often confronted with many possible independent variables from which to choose a "best set". Biological intuition should definitely play a role in the choice. The Partial t/F tests may be used to determine whether the addition of any particular independent variable X_i makes a difference to a model consisting of other X's previously constructed. Statistical software typically provides output to assist in the choice of variables, although they differ in details. Commonly encountered are tables listing "extra" sums of squares for each independent variable with associated F statistics and P values. It is important to note that all "extra" sums of squares add to Sum of Squares Regression (SSR) for what is termed the "Full Model" - *a multiple regression including all independent variables*. Each "extra" sum of squares represents the decrease in SSR, or equivalently the increase in Sum of Squares Error (SSE), with subtraction of an independent variable from the multiple regression producing what is termed a "Reduced Model".

Of course, there are multiple ways to subtract an independent variable from a Full Model to produce a Reduced Model. Two of the most common are described here.

Marginal "Extra" Sums of Squares:

In this instance, only a single independent variable is removed at a time from the Full Model (FM) to produce one or more Reduced Models (RM). Resultant comparisons between RM's and FM thus involves statistical hypotheses of the form: H_0: $\beta_i = 0$ versus H_1: $\beta_i \neq 0$ **for each independent variable at a time, *with all other independent variables retained within the Reduced Model.* "Extra" Sums of Squares (and tests of this type) are called "Marginal" or "Type II" Regression Sums of Squares.**

ZAR := READPRN("c:/DATA/Biostatistics/ZarEX20.1aR.txt") **< Zar Example 20.1a**

Prototype in R:

```
#MULTIPLE REGRESSION MODEL SS
#ZAR EXAMPLE 20.1a
ZAR=read.table("c:/DATA/Biostatistics/ZarEX20.1aR.txt")
ZAR
attach(ZAR)

Y=var5
X1=var1
X2=var2
X3=var3
X4=var4

#DOWNLOAD {CAR} PACKAGE FROM CRAN:
library(car)
#TYPE 2 SS:

FM=lm(Y~X1+X2+X3+X4)
```

```
summary(FM)
Anova(FM)
```

```
> summary(FM)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.5070 -0.1112  0.0401  0.1550  0.9617 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.95828    1.36361   2.169  0.03869 *  
X1          -0.12932    0.02129  -6.075 1.50e-06 ***
X2          -0.01878    0.05628  -0.334  0.74102    
X3          -0.04621    0.20727  -0.223  0.82518    
X4           0.20876    0.06703   3.114  0.00423 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4238 on 28 degrees of freedom
Multiple R-squared: 0.6589,     Adjusted R-squared: 0.6102 
F-statistic: 13.52 on 4 and 28 DF,  p-value: 2.948e-06
```

Partial t-statistics >

Here the Anova() function in package {car} produces marginal sums of squares, F statistics, and P values that are identical to the Partial t/F tests shown in Lecture Worksheet 380 (although only t values are given by R's summary() function above). These Partial tests are "marginal" in the sense that only one independent variable is removed from FM at a time for each test and all other variables are left within the RM.

$6.075^2 = 36.9056$ **< t^2 statistic for X_1**

^ note t^2 above = F here

```
> Anova(FM)
Anova Table (Type II tests)

Response: Y
          Sum Sq Df F value    Pr(>F)    
X1        6.6298  1 36.9063 1.496e-06 ***
X2        0.0200  1  0.1114  0.741024    
X3        0.0089  1  0.0497  0.825178    
X4        1.7421  1  9.6979  0.004227 ** 
Residuals 5.0299 28                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Partial F statistics >

```
#REDUCED MODELS:
RM1=lm(Y~X2+X3+X4)  #subtracting X1
RM2=lm(Y~X1+X3+X4)  #subtracting X2
RM3=lm(Y~X1+X2+X4)  #subtracting X3
RM4=lm(Y~X1+X2+X3)  #subtracting X4

Anova(RM1)
Anova(RM2)
Anova(RM3)
Anova(RM4)
```

```
> Anova(RM1)
Anova Table (Type II tests)

Response: Y
           Sum Sq Df F value  Pr(>F)  
X2         1.3777  1  3.4267 0.07436 .
X3         0.5477  1  1.3622 0.25267  
X4         2.3125  1  5.7516 0.02313 *
Residuals 11.6596 29                  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note the *change* in Residual SS between FM above and Reduced Model RM_1 excluding independent variable X_1. This difference is shown as the "Extra" SS in Anova(FM) above.

Serial "Extra" Sums of Squares:

"Serial" Sums of Squares, also known as "Type I" SS and associated Partial F-tests represent the *change* in SSR (or SSE) when independent variables are added or deleted in an ordered fashion. In R, the order is determined by the order given to independent variables in the formula within the lm() function from left to right. Thus, at each step of addition or deletion of an independent variable, the Full Model (FM) and (RM) changes in a serial fashion.

```
TYPE 1 SS:
anova(FM)

FM1=lm(Y~X1+X2)
RM1=lm(Y~X1)
anova(FM1)
anova(RM1)
```

^ Here serial addition of independent variables follows the order seen in the lm() function: X_1,X_2,X_3,X_4.

anova(FM) shows the change in SSR (or SSE) with the *addition* of X_1, then X_2, followed by X_3, and then X_4. Alternatively it shows the *subtraction* of X_4 first followed by X_3, X_2 and then X_1.

```
> anova(FM)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)    
X1         1 7.8762  7.8762 43.8450 3.496e-07 ***
X2         1 0.0131  0.0131  0.0732  0.788785    
X3         1 0.0859  0.0859  0.4781  0.494963    
X4         1 1.7421  1.7421  9.6979  0.004227 ** 
Residuals 28 5.0299  0.1796                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

anova(FM_1) shows an intermediate step in which only independent variables X_1 & X_2 have been added in order.

```
> anova(FM1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)    
X1         1 7.8762  7.8762 34.4549 2.007e-06 ***
X2         1 0.0131  0.0131  0.0575    0.8121    
Residuals 30 6.8579  0.2286                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

anova(RM_1) shows an intermediate step in which only independent variable X_1 has been added.

```
> anova(RM1)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)    
X1         1 7.8762  7.8762  35.535 1.371e-06 ***
Residuals 31 6.8710  0.2216                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```