ORIGIN ≡ 0

Regression Testing & Extra Sums of Squares

W. Stein

In Multiple regression, multiple tests are available to determine the potential value of including particular independent variables in a Linear model fit. Terminology is nicely summarized in Section 7.3 of Kuter et al. (KNNL) *Applied Linear Statistical Models* 5th Edition. Standard statistical packages present results of these tests directly, but often in a way that requires decipherment. Below is an analysis of the R statistical package. For information on how to fit a multiple regression model, see KNNL Chapter 6 & 2008 Linear Models 07.

Example:

Continuing our example from 2008 Linear Models 07, data comes from R's ISwR package as dataset "cystfibr".

Prototype in R:

```
#LOAD PACKAGES
require(ISwR)
require(car)
#LOAD DATA FROM ISwR
data(cystfibr)
attach(cystfibr)
cystfibr
```

```
> cystfibr
   age sex height weight bmp fev1  rv frc tlc pemax
1    7   0    109   13.1  68   32 258 183 137    95
2    7   1    112   12.9  65   19 449 245 134    85
3    8   0    124   14.1  64   22 441 268 147   100
4    8   1    125   16.2  67   41 234 146 124    85
5    8   0    127   21.5  93   52 202 131 104    95
6    9   0    130   17.5  68   44 308 155 118    80
7   11   1    139   30.7  89   28 305 179 119    65
8   12   1    150   28.4  69   18 369 198 103   110
9   12   0    146   25.1  67   24 312 194 128    70
10  13   1    155   31.5  68   23 413 225 136    95
11  13   0    156   39.9  89   39 206 142  95   110
12  14   1    153   42.1  90   26 253 191 121    90
13  14   0    160   45.6  93   45 174 139 108   100
14  15   1    158   51.2  93   45 158 124  90    80
15  16   1    160   35.9  66   31 302 133 101   134
16  17   1    153   34.8  70   29 204 118 120   134
17  17   0    174   44.7  70   49 187 104 103   165
18  17   1    176   60.1  92   29 188 129 130   120
19  17   0    171   42.6  69   38 172 130 103   130
20  19   1    156   37.2  72   21 216 119  81    85
21  19   0    174   54.6  86   37 184 118 101    85
22  20   0    178   64.0  86   34 225 148 135   160
23  23   0    180   73.8  97   57 171 108  98   165
24  23   0    175   51.1  71   33 224 131 113    95
25  23   0    179   71.5  95   52 225 127 101   195
```

Fitting Full Model:

```
#FITTING FULL MODEL
FM=lm(pemax~age+sex+height+weight
+bmp+fev1+rv+frc+tlc)
FM
```

Reporting Coefficients:

```
> FM
Call:
lm(formula = pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc + tlc)
Coefficients:
(Intercept)          age          sex       height       weight          bmp
   176.0582      -2.5420      -3.7368      -0.4463       2.9928      -1.7449
       fev1           rv          frc          tlc
     1.0807       0.1970      -0.3084       0.1886
```

Overall F test and *Partial* t/F tests:

```
#OVERALL F TEST & PARTIAL t/F TESTS
summary(FM)
```

```
> summary(FM)
Call:
lm(formula = pemax ~ age + sex + height + weight + bmp +
fev1 +
    rv + frc + tlc)
Residuals:
    Min      1Q  Median      3Q     Max
-37.338 -11.532   1.081  13.386  33.405

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 176.0582   225.8912   0.779    0.448
age          -2.5420     4.8017  -0.529    0.604
sex          -3.7368    15.4598  -0.242    0.812
height       -0.4463     0.9034  -0.494    0.628
weight        2.9928     2.0080   1.490    0.157
bmp          -1.7449     1.1552  -1.510    0.152
fev1          1.0807     1.0809   1.000    0.333
rv            0.1970     0.1962   1.004    0.331
frc          -0.3084     0.4924  -0.626    0.540
tlc           0.1886     0.4997   0.377    0.711

Residual standard error: 25.47 on 15 degrees of freedom
Multiple R-squared: 0.6373,     Adjusted R-squared: 0.4197
F-statistic: 2.929 on 9 and 15 DF,  p-value: 0.03195
```

< Partial t tests for each β reported by R

^ Overall F test for all β's - See F_T under Type I anova{stats} below

Calculation of Partial F-tests from Type II Anova{car}

```
#Anova{car} SET FOR TYPE II SS
#RETURNS REGRESSION SS DIFFERENCE FOR EACH VAR
#AND RESIDUAL SS WHEN ALL VARS ARE INCLUDED
#RESIDUAL SS SAME AS IN TYPE I
Anova(FM,type="2")
```

Anova() in the {car} package provides the same *Partial* test using F statistic, as can be verified by squarring each value of the t statistic above. As expected, probability values are identical. >

```
> Anova(FM,type="2")
Anova Table (Type II tests)

Response: pemax
          Sum Sq Df F value Pr(>F)
age        181.8  1  0.2803 0.6043
sex         37.9  1  0.0584 0.8123
height     158.3  1  0.2440 0.6285
weight    1441.2  1  2.2215 0.1568
bmp       1480.1  1  2.2815 0.1517
fev1       648.4  1  0.9995 0.3333
rv         653.8  1  1.0077 0.3314
frc        254.6  1  0.3924 0.5405
tlc         92.4  1  0.1424 0.7112
Residuals 9731.2 15
```

Note: as described in R's documentation for Anova{car}, Anova Types II & III are derived from similar terminology in SAS, although with some disagreement about exact usage (involving "higher-order" terms in the models). Anova Types II/III are "marginal" tests in which a difference in Sums of Squares Error SSE (or equivalently difference in Sums of Squares Regression SSR) is calculated for a single independent variable when compared with all other independent variables *already* within the regression model.

$$Fp := \frac{\frac{(1480.1)}{1}}{\frac{9731.4}{15}}$$

< partial SSR for bmp

< MSE for FM

$Fp = 2.2814$ **< same as in Anova{car} above. Also, equivalent to t-test in summary(FM) above**

GLM F test for Full versus Reduced Model:

```
#FITTING REDUCED MODEL - ONLY "age","sex","weight" ,"bmp"
RM=lm(pemax~age+sex+weight+bmp)
#GLM F TEST of RM vs FM
anova(RM,FM)
```

```
> anova(RM,FM)
Analysis of Variance Table
Model 1: pemax ~ age + sex + weight + bmp
Model 2: pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc + tlc

                     Res.Df     RSS Df Sum of Sq     F Pr(>F)
Reduced Model SSE >  1     20 12602.1
Full Model SSE    >  2     15  9731.2  5    2870.9 0.885 0.5149
                                                   ^ GLM F test
                                         ^ "Extra" Sum of Squares
```

Calculation of GLM F-test from function anova{stats}

$$F_{GLM} := \frac{\frac{2870.9}{20-15}}{\frac{9731.2}{15}}$$

< $SSE_{Reduced\ Model}$ - $SSE_{Full\ Model}$
< $df_{Reduced\ Model}$ - $df_{Full\ Model}$
< $MSE_{Full\ Model}$

$F_{GLM} = 0.8851$ **< same as above**

Note: this test is quite general. Any combination of independent variables may be tested against any other by specifying a comparison between FM and RM. Here the difference between Residual (Error) Sum of Squares SSE for the Full model and for the Reduced model is reported by R:anova. This is the "extra" sum of squares for the test.

Extra Sums of Squares:

Multiple "Extra" Sums of Squares, all derived from a difference in SSE (or SSR) between Full models (FM) versus Reduced models (RM), are usually calculated by statistical packages and form the basis for F tests of the GLM type as above. In general, F tests can be described as "marginal" tests if an RM excludes *only one* independent variable from the FM. In this case, an "Extra" Sum of Squares measures the change in SSE/SSR directly attributed to the excluded variable with all other independent variables already fitted in a linear model. By contrast, other F tests can be described as "serial" in which independent variables are added one at a time into a progressively more complex linear model. With each step of addition into a growing linear model, both FM and RM may exclude some, or even most, independent variables. In this case, "Extra" Sums of Squares, as calculated above, only measure differences in SSE (or SSR) for partial RM & FM models as specified by each step. Order of addition of variables is also important because different orders of entry to different growing linear models typically produce different partitions of SSE (or SSR). In R, as in most statistical packages, order of entry into a "serial" linear model is specified by the order given in the formula from left to right.

Calculation of Overall F-test from Type I anova{stats}

```
#anova{stats} = TYPE I SS
#RETURNS DIFFERENCE IN REGRESSION SS
#AT TIME OF ENTRY INTO THE MODEL
 IN ORDER FROM TOP TO BOTTOM
anova(FM)
```

Note that anova{stats} reports "serial" or Type I "Extra" Sums of Squares & F tests using a GLM type tests between RM & FM sequentially as specified in the formula in lm(FM) as read from left to right. To see how this is calculated, anova() can be used for models at each step - see below.

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

Response: pemax
          Df  Sum Sq Mean Sq F value   Pr(>F)
age        1 10098.5 10098.5 15.5661 0.001296 **
sex        1   955.4   955.4  1.4727 0.243680
height     1   155.0   155.0  0.2389 0.632089
weight     1   632.3   632.3  0.9747 0.339170
bmp        1  2862.2  2862.2  4.4119 0.053010 .
fev1       1  1549.1  1549.1  2.3878 0.143120
rv         1   561.9   561.9  0.8662 0.366757
frc        1   194.6   194.6  0.2999 0.592007
tlc        1    92.4    92.4  0.1424 0.711160
Residuals 15  9731.2   648.7
```

$$F_T := \frac{\dfrac{(10098.5+955.4+155.0+632.3+2862.2+1549.1+561.9+194.6+92.4)}{9}}{\dfrac{9731.1}{15}}$$

< add serial SSR's & divide by Σdf=1 for each

< MSE for FM

$F_T = 2.929$ **< same as in summary(FM) above.**

$$\frac{(15.5661 + 1.4727 + 0.2389 + 0.9747 + 4.4119 + 2.3878 + 0.8662 + 0.2999 + 0.1424)}{9} = 2.929$$

^ alternate calculation of F_T from F-test values in anova(FM)

Calculation of Partial F-tests from Type I anova{stats}

$$F_S := \frac{\dfrac{(92.4+194.6+561.9)}{3}}{\dfrac{9731.1}{15}}$$

< Σ partial SSR's for tlc,frc,rv (working in reverse order of entry)
< Σ partial df's

<MSE for FM

$F_S = 0.4362$ **< same result as in anova{stats} below**

$$\frac{0.1424 + 0.2999 + 0.8662}{3} = 0.4362$$

^ alternate calculation of F_S from F-test values in anova(FM)

Comparing RM & FM from anova{stats}

```
#COMPARING RM & FM FOR PARTIAL F-TEST
RM=lm(pemax~age+sex+height+weight+bmp+fev1)
FM=lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
anova(RM,FM)
```

$$\frac{\dfrac{848.9}{(18-15)}}{\dfrac{9731.2}{15}} = 0.4362$$

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

Model 1: pemax ~ age + sex + height + weight + bmp + fev1
Model 2: pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc +
    tlc
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     18 10580.1
2     15  9731.2  3     848.9 0.4362 0.7303
```

^ same as F_S

Note: this demonstrates the general usefulness of using Type I ANOVA, as long as one loads FM in a proper order. From this setup, one can easily test last entered independent variables to determine whether they are acturally needed, or not.

Comparing serial anova{stats} with GLM's RM & FM at one step in "serial" model construction:

```
#COMPARING RM & FM UPON ENTRY OF "fev1"
FM=lm(pemax~age+sex+height+weight+bmp+fev1)
RM=lm(pemax~age+sex+height+weight+bmp)
anova(RM,FM)
```

< note *limited* FM (missing rv, frc, tlc)

"Extra" Sum of Squares calculated at fev1 step corresponds to entry in the anova{stats} report above. Note also that at this step in "serial" model construction, FM includes some but not all of the independent variables. The F tests based on these RM & FM are thus "partial" tests in a non-marginal sense. If a researcher views the partial RM and *limited* FM models to have meaning, then Type I tests make sense. Otherwise they do not. >

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

Model 1: pemax ~ age + sex + height + weight + bmp
Model 2: pemax ~ age + sex + height + weight +
bmp + fev1
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     19 12129.2
2     18 10580.1  1    1549.1 2.6355 0.1219
```

$$F_{GLM} := \frac{\frac{1549.1}{19-18}}{\frac{10580}{18}} \qquad F_{GLM} = 2.6355$$

```
#EQUIVALENT anova{stats} APPROACH
#INVOLVING A LIMITED FM IGNORING LAST ENTERED VARIABLES
FM=lm(pemax~age+sex+height+weight+bmp+fev1)
anova(FM)
```

Note: Same partial F approach as above:

$$F_S := \frac{\frac{1549.1}{1}}{\frac{10580.1}{18}}$$

< Extra SS for fev1 divided by df=1

< MSE for *limited* FM as defined here and for GLM approach

$F_S = 2.6355$ **< same as F_{GLM} above**

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

Response: pemax
          Df  Sum Sq Mean Sq F value    Pr(>F)
age        1 10098.5 10098.5 17.1806 0.0006083 ***
sex        1   955.4   955.4  1.6255 0.2185424
height     1   155.0   155.0  0.2637 0.6138681
weight     1   632.3   632.3  1.0758 0.3133741
bmp        1  2862.2  2862.2  4.8695 0.0405636 *
fev1       1  1549.1  1549.1  2.6355 0.1218847
Residuals 18 10580.1   587.8
```

Automated Comparisons in R: drop1{stats}

```
#RECOVERY OF TYPE II SS USING drop1{stats}
#RSS RETURNS RESIDUAL SS WHEN SPECIFIC VAR IS EXCLUDED
FM=lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
drop1(FM)
```

drop1{stats} provides an alternate way to calculate "marginal" or Type II SS and F tests by taking FM and dropping each independent variable at a time. Each comparision involves the complete FM, but different RM's one for each variable exluded. > **Compare with the GLM type result below. For definition of AIC, see KNNL Ch.9.**

```
> drop1(FM)
Single term deletions
Model:
pemax ~ age + sex + height + weight + bmp + fev1 + rv +
frc + tlc
       Df Sum of Sq     RSS   AIC
<none>               9731.2 169.1
age     1     181.8  9913.1 167.6
sex     1      37.9  9769.2 167.2
height  1     158.3  9889.6 167.5
weight  1    1441.2 11172.5 170.6
bmp     1    1480.1 11211.4 170.6
fev1    1     648.4 10379.7 168.7
rv      1     653.8 10385.0 168.7
frc     1     254.6  9985.8 167.8
tlc     1      92.4  9823.7 167.3
```

Comparing "marginal" deletion of one variable:

```
#CREATING RM - EXCLUDING ONLY "rv"
RM=lm(pemax~age+sex+height+weight+bmp+fev1+frc+tlc)
anova(RM,FM)
```

SSE and "Extra" SS here match report > from drop1{stats}.

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

Model 1: pemax ~ age + sex + height + weight + bmp + fev1 + frc + tlc
Model 2: pemax ~ age + sex + height + weight + bmp + fev1 + rv + frc + tlc
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     16 10385.0
2     15  9731.2  1     653.8 1.0077 0.3314
```

Automated Comparisons in R: add1{stats}

```
#CREATING RM - WITH ONLY "age"
RM=lm(pemax~age)
#MARGINAL COMPARISION BY ADDING EACH VARIABLE IN TURN
add1(RM,~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
```

add1{stats} calculates "Extra" SS and Residual or Error SSE upon adding a single independent variable above what is already in RM. Compare with GLM Result below

```
> add1(RM,~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
Single term additions

Model:
pemax ~ age
       Df Sum of Sq     RSS    AIC
<none>              16734.2  166.7
sex     1     955.4 15778.7  167.2
height  1     182.3 16551.8  168.4
weight  1     945.2 15789.0  167.2
bmp     1       0.2 16734.0  168.7
fev1    1    2185.1 14549.0  165.2
rv      1      20.5 16713.7  168.6
frc     1      28.3 16705.8  168.6
tlc     1     389.1 16345.0  168.1
```

Comparing "marginal" entry of one variable:

```
#COMPARING SINGLE ADDITION OF "fev1"
FM=lm(pemax~age+fev1)
anova(RM,FM)
```

Same result for SSE & "Extra" SS considering only variable "fev1" >

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

Model 1: pemax ~ age
Model 2: pemax ~ age + fev1
  Res.Df     RSS Df Sum of Sq      F  Pr(>F)
1     23 16734.2
2     22 14549.0  1    2185.1 3.3042 0.08275 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Automated Simple Regression SS:

```
#SIMPLE REGRESSIONS WITH add1{stats}
#NOTE IN MODEL 1=include constant, -1=no constant
RM=lm(pemax~1)
add1(RM,~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
```

```
> add1(RM,~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
Single term additions

Model:
pemax ~ 1
       Df Sum of Sq     RSS    AIC
<none>              26832.6  176.5
age     1   10098.5 16734.2  166.7
sex     1    2234.4 24598.2  176.3
height  1    9634.6 17198.0  167.3
weight  1   10827.2 16005.5  165.5
bmp     1    1413.5 25419.2  177.1
fev1    1    5515.4 21317.2  172.7
rv      1    2671.8 24160.9  175.8
frc     1    4670.6 22162.1  173.7
tlc     1     885.1 25947.6  177.6
```

```
#SIMPLE REGRESSION FOR WEIGHT ONLY
anova(lm(pemax~weight))
```

```
> anova(lm(pemax~weight))
Analysis of Variance Table

Response: pemax
          Df  Sum Sq Mean Sq F value    Pr(>F)
weight     1 10827.2 10827.2  15.559 0.0006457 ***
Residuals 23 16005.5   695.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Automated Stepwise Regressions:

```
#AUTOMATED STEPWISE REGRESSION USING AIC
FM=lm(pemax~age+sex+height+weight+bmp+fev1+rv+frc+tlc)
RM=lm(pemax~1)

# LOAD R PACKAGE {MASS}
require(MASS)
stepAIC(FM,direction="backward")

stepAIC(RM,~age+sex+height+weight+bmp+fev1+rv+frc+tlc,direction="forward")
```