ORIGIN ≡ 0

Higher Order Models in Multiple Regression

W. Stein

So called "higher-order" linear models in multiple regression involve inclusion of polynomial factors (squares, cubes, etc. of an independent variable) as well as interaction terms (two or more independent variables multiplied). Polynomial factors are useful in modeling non-linear relationships especially when such "higher-order" factors make some interpretive sense. Otherwise, a transformation of one kind or another remains a viable alternative. Interaction terms, by contrast, are very useful in testing for non-additive effects of independent variables on the response variable. In general, once "higher-order" variables are constructed, entry into a multiple regression is straight forward. Example below comes from Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition.

Strangely Scaled Polynomial Example:

Power Cells Example KNNL Table 8.1

Reading Data using R:

```
#POLYNOMIAL REGRESSION
# READ STRUCTURED DATA TABLE
K=read.table("c:/2008LinearModelsData/PowerCellsR.txt")
K
attach(K)
options(digits=10)
```

```
> K
     Y  X1 X2
1  150 0.6 10
2   86 1.0 10
3   49 1.4 10
4  288 0.6 20
5  157 1.0 20
6  131 1.0 20
7  184 1.0 20
8  109 1.4 20
9  279 0.6 30
10 235 1.0 30
11 224 1.4 30
```

Scaling X values using R:

```
#SCALE AND SQUARE/MULTIPLY X DATA ONLY
x1=scale(X1,center=T,scale=0.4)
x2=scale(X2,center=T,scale=10)
x1s=x1^2
x2s=x2^2
x1x2=x1*x2

#REFORMAT AS DATAFRAME
Kscaled=data.frame(Y,x1,x2,x1s,x2s,x1x2)
Kscaled
detach(K)
attach(Kscaled)
```

confirmed KNNL p. 301 >

```
> Kscaled
     Y x1 x2 x1s x2s x1x2
1  150 -1 -1   1   1    1
2   86  0 -1   0   1    0
3   49  1 -1   1   1   -1
4  288 -1  0   1   0    0
5  157  0  0   0   0    0
6  131  0  0   0   0    0
7  184  0  0   0   0    0
8  109  1  0   1   0    0
9  279 -1  1   1   1   -1
10 235  0  1   0   1    0
11 224  1  1   1   1    1
```

Fitting the Polynomial Linear Model in R:

```
#FITTING FULL LINEAR MODEL
FM=lm(Y~x1+x2+x1s+x2s+x1x2)
summary(FM)
```

coefficients verified KNNL p. 301>

```
> summary(FM)
Call: lm(formula = Y ~ x1 + x2 + x1s + x2s + x1x2)
Residuals:
         1          2          3          4          5          6
-21.464912   9.263158  12.201754  41.929825  -5.842105 -31.842105
         7          8          9         10         11
 21.157895 -25.403509 -20.464912   7.263158  13.201754
Coefficients:
             Estimate Std. Error  t value   Pr(>|t|)
(Intercept) 162.84211   16.60761  9.80527 0.00018784 ***
x1          -55.83333   13.21670 -4.22445 0.00829229 **
x2           75.50000   13.21670  5.71247 0.00229727 **
x1s          27.39474   20.34008  1.34684 0.23585632
x2s         -10.60526   20.34008 -0.52140 0.62435225
x1x2         11.50000   16.18709  0.71044 0.50918373
```

Partial F-Test for higher-order terms using R:

```
#PARTIAL F TEST FOR HIGHER ORDER VARIABLES
anova(FM)
F=((529+285+1646)/3)/1048
F
qf(0.95,3,5)
```

The F-test involves the ratio:
Numerator = "Extra" SSR for terms on entry/df
Denominator = SSE for Full Model/df

These were calculated from the SSE & df columns qf() gives the appropriate F-distribution Critical Value.

Results verified KNNL p. 304

Null hypothesis is NOT rejected meaning the higher-order variables are NOT necessary...

```
> anova(FM)
Analysis of Variance Table
Response: Y
          Df    Sum Sq   Mean Sq  F value    Pr(>F)
x1         1 18704.167 18704.167 17.84599 0.0082923 **
x2         1 34201.500 34201.500 32.63229 0.0022973 **
x1s        1  1645.967  1645.967  1.57045 0.2655520
x2s        1   284.928   284.928  0.27186 0.6243522
x1x2       1   529.000   529.000  0.50473 0.5091837
Residuals  5  5240.439  1048.088
> F=((529+284.928+1645.967)/3)/1048.088
> F
[1] 0.7823436582
> qf(0.95,3,5)
[1] 5.409451318
```

< values verified KNNL p. 304

529.000 + 284.928 + 1645.967 = 2459.895

GLM Test for higher-order terms in R:

```
#GLM TEST FOR HIGHER ORDER VARIABLES
FM=lm(Y~x1+x2+x1s+x2s+x1x2)
RM=lm(Y~x1+x2)
anova(RM,FM)
```

Equivalence of test confirmed...

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

Model 1: Y ~ x1 + x2
Model 2: Y ~ x1 + x2 + x1s + x2s + x1x2
  Res.Df       RSS Df Sum of Sq       F Pr(>F)
1      8 7700.3333
2      5 5240.4386  3 2459.8947 0.78234 0.5527
```

Interactions Example:

Body Fat Example KNNL Table 7.1

Reading Data using R:

```
#INTERACTIONS REGRESSION
# READ A STRUCTURED DATA TABLE
K=read.table("c:/2008LinearModelsData/BodyFatR.txt")
K
attach(K)
```

```
> K
      X1   X2   X3    Y
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
7  31.4 58.5 27.6 27.1
8  27.9 52.1 30.6 25.4
9  22.1 49.9 23.2 21.3
10 25.5 53.5 24.8 19.3
11 31.1 56.6 30.0 25.4
12 30.4 56.7 28.3 27.2
13 18.7 46.5 23.0 11.7
14 19.7 44.2 28.6 17.8
15 14.6 42.7 21.3 12.8
16 29.5 54.4 30.1 23.9
17 27.7 55.3 25.7 22.6
18 30.2 58.6 24.6 25.4
19 22.7 48.2 27.1 14.8
20 25.2 51.0 27.5 21.1
```

Centering X Values using R:

```
#CENTERING X VALUES
X=cbind(X1,X2,X3)
Xc=data.frame(scale(X,center=T,scale=FALSE))
Xc
attach(Xc)
```

```
> Xc
          X1    X2    X3
1   -5.805 -8.07  1.48
2   -0.605 -1.37  0.58
3    5.395  0.73  9.38
4    4.495  3.13  3.48
5   -6.205 -8.97  3.28
6    0.295  2.73 -3.92
7    6.095  7.33 -0.02
8    2.595  0.93  2.98
9   -3.205 -1.27 -4.42
10   0.195  2.33 -2.82
11   5.795  5.43  2.38
12   5.095  5.53  0.68
13  -6.605 -4.67 -4.62
14  -5.605 -6.97  0.98
15 -10.705 -8.47 -6.32
16   4.195  3.23  2.48
17   2.395  4.13 -1.92
18   4.895  7.43 -3.02
19  -2.605 -2.97 -0.52
20  -0.105 -0.17 -0.12
```

Making Interaction Variables in R:

```
#MAKING INTERACTION VARIABLES
X1X2=X1*X2
X1X3=X1*X3
X2X3=X2*X3
```

Fitting the Full Linear Model with Interactions in R:

```
#FIT OF THE FULL LINEAR MODEL AND REPORT
FM=lm(Y~X1+X2+X3+X1X2+X1X3+X2X3)
summary(FM)
```

```
> summary(FM)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X1X2 + X1X3 + X2X3)

Residuals:
       Min         1Q     Median         3Q        Max
-3.9266581 -0.8728168  0.1681908  1.2929007  3.3309548

Coefficients:
                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) 20.526893531  1.073626456 19.11921 6.6998e-11 ***
X1           3.437808068  3.578665718  0.96064    0.35426
X2          -2.094717339  3.036769568 -0.68978    0.50246
X3          -1.616337237  1.907210684 -0.84749    0.41205
X1X2         0.008875562  0.030850456  0.28770    0.77811
X1X3        -0.084790836  0.073417742 -1.15491    0.26892
X2X3         0.090415385  0.092001297  0.98276    0.34366
```

coefficients verified KNNL p. 312>

Partial F-Test for higher-order terms using R:

```
#PARTIAL F TEST FOR
HIGHER ORDER VARIABLES
anova(FM)
F=((6.51+2.70+1.50)/3)/6.75
F
qf(0.95,3,13)
```

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

Response: Y
          Df    Sum Sq   Mean Sq  F value     Pr(>F)
X1         1 352.26980 352.26980 52.22383 6.6822e-06 ***
X2         1  33.16891  33.16891  4.91728   0.045026 *
X3         1  11.54590  11.54590  1.71167   0.213428
X1X2       1   1.49572   1.49572  0.22174   0.645521
X1X3       1   2.70433   2.70433  0.40092   0.537597
X2X3       1   6.51484   6.51484  0.96582   0.343662
Residuals 13  87.69000   6.74538

> F=((6.51484+2.70433+1.49572)/3)/6.74538
> F
[1] 0.5294927788
> qf(0.95,3,13)
1] 3.410533645
```

values verified KNNL p. 313 >

$6.51484 + 2.70433 + 1.49572 = 10.7149$

GLM Test for higher-order terms in R:

```
#GLM TEST FOR HIGHER ORDER VARIABLES
FM=lm(Y~X1+X2+X3+X1X2+X1X3+X2X3)
RM=lm(Y~X1+X2+X3)
anova(RM,FM)
```

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

Model 1: Y ~ X1 + X2 + X3
Model 2: Y ~ X1 + X2 + X3 + X1X2 + X1X3 + X2X3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 98.405
2     13 87.690  3    10.715 0.5295 0.6699
```

Again, I think rounding is the only difference here...

lm() & glm() Formula Options in R:

#USING FORMULA OPTIONS IN R:
#FULLY CROSSED MODEL WITH ALL HIGHER-ORDER VARIABLES
FM=lm(Y~X1*X2*X3)
summary(Fm)
anova(FM)

> summary(FM)

```
Call: lm(formula = Y ~ X1 * X2 * X3)
Residuals:
    Min      1Q  Median      3Q     Max
-4.0586 -1.2222  0.2303  1.0287  3.3960
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.968892   1.369670  15.309 3.08e-09 ***
X1           2.632790   3.964805   0.664    0.519
X2          -1.423062   3.356662  -0.424    0.679
X3          -1.040456   2.227418  -0.467    0.649
X1:X2       -0.004416   0.040008  -0.110    0.914
X1:X3       -0.118089   0.097105  -1.216    0.247
X2:X3        0.095440   0.095042   1.004    0.335
X1:X2:X3    -0.006789   0.012455  -0.545    0.596
```

> anova(FM)

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 49.4004 1.378e-05 ***
X2         1  33.17   33.17  4.6514   0.05202 .
X3         1  11.55   11.55  1.6191   0.22732
X1:X2      1   1.50    1.50  0.2098   0.65514
X1:X3      1   2.70    2.70  0.3792   0.54951
X2:X3      1   6.51    6.51  0.9136   0.35802
X1:X2:X3   1   2.12    2.12  0.2972   0.59566
Residuals 12  85.57    7.13
```

^ Note additional X1:X2:X3 term here ^

#SPECIFYING HIGHER-ORDER VARIABLES
FM=lm(Y~X1,X2,X3,X1:X2,X1:X3,X2:X3)
anova(FM)

same as above >

> summary(FM)

```
Call: lm(formula = Y ~ X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3)
Residuals:
    Min      1Q  Median      3Q     Max
-3.9267 -0.8728  0.1682  1.2929  3.3310
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 20.526894   1.073626  19.119  6.7e-11 ***
X1           3.437808   3.578666   0.961    0.354
X2          -2.094717   3.036770  -0.690    0.502
X3          -1.616337   1.907211  -0.847    0.412
X1:X2        0.008876   0.030850   0.288    0.778
X1:X3       -0.084791   0.073418  -1.155    0.269
X2:X3        0.090415   0.092001   0.983    0.344
```

same as above >

```
>anova(FM)
Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 352.27  352.27 52.2238 6.682e-06 ***
X2         1  33.17   33.17  4.9173   0.04503 *
X3         1  11.55   11.55  1.7117   0.21343
X1:X2      1   1.50    1.50  0.2217   0.64552
X1:X3      1   2.70    2.70  0.4009   0.53760
X2:X3      1   6.51    6.51  0.9658   0.34366
Residuals 13  87.69    6.75
```