ORIGIN ≡ 0

Standardizing Data in Multiple Regression

W. Stein

In order to compare regression coefficients in multiple linear regression, variables are often "standardized" so that means of each are zero and variance is the same for each, typically 1. Example below comes from Kuter et al. (KNNL) A*pplied Linear Statistical Models* 5th Edition.

Example:

Dwaine Studios Example KNNL Table 6.5 & Table 7.5

K := READPRN("c:/2008LinearModelsData/Dwaine.txt")

$X1 := K^{\langle 0 \rangle}$ $X2 := K^{\langle 1 \rangle}$ $Y := K^{\langle 2 \rangle}$

$n := \text{length}\left(K^{\langle 0 \rangle}\right)$ $n = 21$

$$K = \begin{pmatrix} 68.5 & 16.7 & 174.4 \\ 45.2 & 16.8 & 164.4 \\ 91.3 & 18.2 & 244.2 \\ 47.8 & 16.3 & 154.6 \\ 46.9 & 17.3 & 181.6 \\ 66.1 & 18.2 & 207.5 \\ 49.5 & 15.9 & 152.8 \\ 52 & 17.2 & 163.2 \\ 48.9 & 16.6 & 145.4 \\ 38.4 & 16 & 137.2 \\ 87.9 & 18.3 & 241.9 \\ 72.8 & 17.1 & 191.1 \\ 88.4 & 17.4 & 232 \\ 42.9 & 15.8 & 145.3 \\ 52.5 & 17.8 & 161.1 \\ 85.7 & 18.4 & 209.7 \\ 41.3 & 16.5 & 146.4 \\ 51.7 & 16.3 & 144 \\ 89.6 & 18.1 & 232.6 \\ 82.7 & 19.1 & 224.1 \\ 52.3 & 16 & 166.5 \end{pmatrix}$$

Summary Statistics:

Means:

$X1_{bar} := \text{mean}(X1)$ $X1_{bar} = 62.019$

$X2_{bar} := \text{mean}(X2)$ $X2_{bar} = 17.1429$

$Y_{bar} := \text{mean}(Y)$ $Y_{bar} = 181.9048$

< all values confirmed KNNL p. 277

Population Standard Deviations:

$s_{X1} := \text{Stdev}(X1)$ $s_{X1} = 18.6203$

$s_{X2} := \text{Stdev}(X2)$ $s_{X2} = 0.9703$

$s_Y := \text{Stdev}(Y)$ $s_Y = 36.1913$

Multiple Regression using R:

```
# READ A STRUCTURED DATA TABLE
K=read.table("c:/2008LinearModelsData/DwaineDataFrame.txt")
K
attach(K)
#FIT LINEAR MODEL & REPORT
FM=lm(Y~X1+X2)
summary(FM)
```

Regression confirmed KNNL p. 237 >

```
> summary(FM)
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
     Min       1Q   Median       3Q      Max
-18.4239  -6.2161   0.7449   9.4356  20.2151

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -68.8571    60.0170  -1.147   0.2663
X1            1.4546     0.2118   6.868    2e-06 ***
X2            9.3655     4.0640   2.305   0.0333 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 11.01 on 18 degrees of freedom
Multiple R-squared: 0.9167,     Adjusted R-squared:
0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

Standardization & the Correlation Transformation:

$i := 0 .. n - 1$ **< index variable**

Standardization:

$$X1_{s_i} := \frac{X1_i - X1_{bar}}{s_{X1}} \qquad X2_{s_i} := \frac{X2_i - X2_{bar}}{s_{X2}} \qquad Y_{s_i} := \frac{Y_i - Y_{bar}}{s_Y}$$

$K_S := \text{augment}(X1_S, X2_S, Y_S)$

Correlation Transformation:

$$K_c := \frac{1}{\sqrt{n-1}} \cdot K_S$$

$$K_c = \begin{pmatrix} 0.07783 & -0.10205 & -0.04637 \\ -0.20198 & -0.07901 & -0.10815 \\ 0.35163 & 0.24361 & 0.38489 \\ -0.17075 & -0.19423 & -0.1687 \\ -0.18156 & 0.03621 & -0.00188 \\ 0.04901 & 0.24361 & 0.15814 \\ -0.15034 & -0.2864 & -0.17982 \\ -0.12032 & 0.01317 & -0.11557 \\ -0.15754 & -0.1251 & -0.22554 \\ -0.28364 & -0.26336 & -0.27621 \\ 0.3108 & 0.26665 & 0.37068 \\ 0.12947 & -0.00988 & 0.05681 \\ 0.3168 & 0.05926 & 0.30951 \\ -0.2296 & -0.30945 & -0.22616 \\ -0.11431 & 0.15143 & -0.12854 \\ 0.28438 & 0.2897 & 0.17173 \\ -0.24881 & -0.14814 & -0.21937 \\ -0.12392 & -0.19423 & -0.23419 \\ 0.33121 & 0.22056 & 0.31322 \\ 0.24835 & 0.451 & 0.2607 \\ -0.11671 & -0.26336 & -0.09518 \end{pmatrix}$$

$$K_S = \begin{pmatrix} 0.34806 & -0.45639 & -0.20736 \\ -0.90326 & -0.35333 & -0.48367 \\ 1.57253 & 1.08945 & 1.72128 \\ -0.76363 & -0.86862 & -0.75446 \\ -0.81196 & 0.16195 & -0.00842 \\ 0.21917 & 1.08945 & 0.70722 \\ -0.67233 & -1.28084 & -0.80419 \\ -0.53807 & 0.05889 & -0.51683 \\ -0.70456 & -0.55945 & -1.00866 \\ -1.26845 & -1.17778 & -1.23523 \\ 1.38993 & 1.19251 & 1.65773 \\ 0.57899 & -0.04417 & 0.25407 \\ 1.41678 & 0.265 & 1.38418 \\ -1.02678 & -1.3839 & -1.01142 \\ -0.51122 & 0.67723 & -0.57486 \\ 1.27178 & 1.29556 & 0.76801 \\ -1.11271 & -0.6625 & -0.98103 \\ -0.55418 & -0.86862 & -1.04734 \\ 1.48123 & 0.98639 & 1.40076 \\ 1.11067 & 2.01695 & 1.16589 \\ -0.52196 & -1.17778 & -0.42565 \end{pmatrix}$$

< confirmed KNNL p. 277

Transformation Prototype in R:

```
COMMANDS:
#TRANSFORMATIONS OF DATA
n=length(Y)
Ks=scale(K,center=T,scale=T)
ks
Kc=(1/sqrt(n-1))*Ks
Kc
```

> Ks

```
           X1          X2          Y
1   0.3480579 -0.45639094 -0.20736368
2  -0.9032627 -0.35333492 -0.48367315
3   1.5725261  1.08944935  1.72127642
4  -0.7636303 -0.86861502 -0.75445643
5  -0.8119646  0.16194517 -0.00842086
6   0.2191665  1.08944935  0.70722067
7  -0.6723323 -1.28083910 -0.80419213
8  -0.5380704  0.05888915 -0.51683029
9  -0.7045551 -0.55944696 -1.00866114
10 -1.2684550 -1.17778308 -1.23523491
11  1.3899300  1.19250537  1.65772525
12  0.5789883 -0.04416687  0.25407314
13  1.4167824  0.26500119  1.38417887
14 -1.0267836 -1.38389512 -1.01142424
15 -0.5112180  0.67722527 -0.57485527
16  1.2717795  1.29556139  0.76800875
17 -1.1127112 -0.66250298 -0.98103020
18 -0.5541818 -0.86861502 -1.04734447
19  1.4812281  0.98639333  1.40075744
20  1.1106653  2.01695353  1.16589439
21 -0.5219590 -1.17778308 -0.42564816
attr(,"scaled:center")
       X1        X2         Y
 62.01905  17.14286 181.90476
attr(,"scaled:scale")
       X1        X2         Y
18.620328  0.970346 36.191304
```

> Kc

```
            X1           X2           Y
1   0.07782811 -0.102052117 -0.046367928
2  -0.20197568 -0.079008091 -0.108152604
3   0.35162753  0.243608280  0.384889109
4  -0.17075294 -0.194228224 -0.168701586
5  -0.18156081  0.036212042 -0.001882962
6   0.04900712  0.243608280  0.158139349
7  -0.15033807 -0.286404330 -0.179822828
8  -0.12031620  0.013168015 -0.115566765
9  -0.15754332 -0.125096144 -0.225543488
10 -0.28363515 -0.263360303 -0.276206922
11  0.31079779  0.266652307  0.370678634
12  0.12946572 -0.009876011  0.056812480
13  0.31680217  0.059256068  0.309511805
14 -0.22959579 -0.309448356 -0.226161335
15 -0.11431183  0.151432174 -0.128541547
16  0.28437855  0.289696333  0.171731978
17 -0.24880979 -0.148140170 -0.219365020
18 -0.12391883 -0.194228224 -0.234193343
19  0.33121266  0.220564254  0.313218885
20  0.24835231  0.451004519  0.260701911
21 -0.11671358 -0.263360303 -0.095177822
attr(,"scaled:center")
       X1        X2         Y
 62.01905  17.14286 181.90476
attr(,"scaled:scale")
       X1        X2         Y
18.620328  0.970346 36.191304
```

Multiple Regression of Transformed Variables using R:

```
#FIT TRANSFORMED DATA TO LINEAR MODEL
detach(K)
Kc=data.frame(Kc) #REFORMAT Kc AS A DATAFRAME
attach(Kc)
FMc=lm(Y~X1+X2)
summary(FMc)
```

Regression coefficients confirmed KNNL p. 277 >

> summary(FMc)

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

Residuals:
      Min        1Q    Median        3Q       Max
-0.113831 -0.038406  0.004602  0.058298  0.124898

Coefficients:
             Estimate Std. Error  t value Pr(>|t|)
(Intercept) -1.128e-17  1.484e-02 -7.6e-16   1.0000
X1           7.484e-01  1.090e-01    6.868    2e-06 ***
X2           2.511e-01  1.090e-01    2.305   0.0333 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

Residual standard error: 0.06801 on 18 degrees of freedom
Multiple R-squared: 0.9167,     Adjusted R-squared: 0.9075
F-statistic:  99.1 on 2 and 18 DF,  p-value: 1.921e-10
```

Some useful things to know about Standardizing Data:

Calculating the Mean Vector:

$i := 0 .. n - 1$ **< index variable of size n**

$J_i := 1$ **< column vector of 1's of size n**

$K_{bar} := \frac{1}{n} \cdot K^T \cdot J$ $\quad K_{bar} = \begin{pmatrix} 62.019 \\ 17.1429 \\ 181.9048 \end{pmatrix}$ **< mean vector of column means in K**

$$K = \begin{pmatrix} 68.5 & 16.7 & 174.4 \\ 45.2 & 16.8 & 164.4 \\ 91.3 & 18.2 & 244.2 \\ 47.8 & 16.3 & 154.6 \\ 46.9 & 17.3 & 181.6 \\ 66.1 & 18.2 & 207.5 \\ 49.5 & 15.9 & 152.8 \\ 52 & 17.2 & 163.2 \\ 48.9 & 16.6 & 145.4 \\ 38.4 & 16 & 137.2 \\ 87.9 & 18.3 & 241.9 \\ 72.8 & 17.1 & 191.1 \\ 88.4 & 17.4 & 232 \\ 42.9 & 15.8 & 145.3 \\ 52.5 & 17.8 & 161.1 \\ 85.7 & 18.4 & 209.7 \\ 41.3 & 16.5 & 146.4 \\ 51.7 & 16.3 & 144 \\ 89.6 & 18.1 & 232.6 \\ 82.7 & 19.1 & 224.1 \\ 52.3 & 16 & 166.5 \end{pmatrix}$$

Mean-Centering Data:

$j := 0 .. 2$ **< index variable of size p = number of columns in K**

$K_m^{\langle j \rangle} := K^{\langle j \rangle} - K_{bar_j}$ **< Mean-Centered Data values in columns in K minus mean for each column**

$I := identity(n)$ **< n X n Identity matrix**

$K^T \cdot \left(I - \frac{1}{n} \cdot J \cdot J^T \right) =$

	0	1	2	3
0	6.481	-16.819	29.281	-14.219
1	-0.4429	-0.3429	1.0571	-0.8429
2	-7.5048	-17.5048	62.2952	-27.3048

^ Matrix algebra returns the transpose of K_m

$$K_m = \begin{pmatrix} 6.481 & -0.4429 & -7.5048 \\ -16.819 & -0.3429 & -17.5048 \\ 29.281 & 1.0571 & 62.2952 \\ -14.219 & -0.8429 & -27.3048 \\ -15.119 & 0.1571 & -0.3048 \\ 4.081 & 1.0571 & 25.5952 \\ -12.519 & -1.2429 & -29.1048 \\ -10.019 & 0.0571 & -18.7048 \\ -13.119 & -0.5429 & -36.5048 \\ -23.619 & -1.1429 & -44.7048 \\ 25.881 & 1.1571 & 59.9952 \\ 10.781 & -0.0429 & 9.1952 \\ 26.381 & 0.2571 & 50.0952 \\ -19.119 & -1.3429 & -36.6048 \\ -9.519 & 0.6571 & -20.8048 \\ 23.681 & 1.2571 & 27.7952 \\ -20.719 & -0.6429 & -35.5048 \\ -10.319 & -0.8429 & -37.9048 \\ 27.581 & 0.9571 & 50.6952 \\ 20.681 & 1.9571 & 42.1952 \\ -9.719 & -1.1429 & -15.4048 \end{pmatrix}$$

Variance-Covariance Matrix:

$\frac{1}{n-1} \cdot K_m^T \cdot K_m = \begin{pmatrix} 346.7166 & 14.1166 & 636.5294 \\ 14.1166 & 0.9416 & 29.3518 \\ 636.5294 & 29.3518 & 1309.8105 \end{pmatrix}$ **< Variance-Covariance matrix of K is the Sum of Squares of mean-centered data, scaled by sample size**

$S := \frac{1}{n-1} \cdot K^T \cdot \left(I - \frac{1}{n} \cdot J \cdot J^T \right) \cdot K$ $\quad S = \begin{pmatrix} 346.7166 & 14.1166 & 636.5294 \\ 14.1166 & 0.9416 & 29.3518 \\ 636.5294 & 29.3518 & 1309.8105 \end{pmatrix}$

^ Variance-Covariance matrix calculated directly from K

Standardized Data:

$$K_s^{\langle j \rangle} := \frac{K_m^{\langle j \rangle}}{\sqrt{S_{j,j}}}$$

< Standardized Data is the mean-centered data with each column divided by the variance found along the main diagonal of S

$$K_s = \begin{pmatrix} 0.3480579 & -0.4563909 & -0.2073637 \\ -0.9032627 & -0.3533349 & -0.4836731 \\ 1.5725261 & 1.0894493 & 1.7212764 \\ -0.7636303 & -0.868615 & -0.7544564 \\ -0.8119646 & 0.1619452 & -0.0084209 \\ 0.2191665 & 1.0894493 & 0.7072207 \\ -0.6723323 & -1.2808391 & -0.8041921 \\ -0.5380704 & 0.0588892 & -0.5168303 \\ -0.7045551 & -0.559447 & -1.0086611 \\ -1.268455 & -1.1777831 & -1.2352349 \\ 1.38993 & 1.1925054 & 1.6577252 \\ 0.5789883 & -0.0441669 & 0.2540731 \\ 1.4167824 & 0.2650012 & 1.3841789 \\ -1.0267836 & -1.3838951 & -1.0114242 \\ -0.511218 & 0.6772253 & -0.5748553 \\ 1.2717795 & 1.2955614 & 0.7680088 \\ -1.1127112 & -0.662503 & -0.9810302 \\ -0.5541818 & -0.868615 & -1.0473445 \\ 1.4812281 & 0.9863933 & 1.4007574 \\ 1.1106653 & 2.0169535 & 1.1658944 \\ -0.521959 & -1.1777831 & -0.4256482 \end{pmatrix}$$

Correlation Matrix:

$$R := \frac{1}{n-1} \cdot K_s^T \cdot \left(I - \frac{1}{n} \cdot J \cdot J^T \right) \cdot K_s$$

Correlation matrix of K is the Variance-Covariance matrix of the Standardized data. >

$$R = \begin{pmatrix} 1 & 0.7813 & 0.9446 \\ 0.7813 & 1 & 0.8358 \\ 0.9446 & 0.8358 & 1 \end{pmatrix}$$

Correlation Transformation:

KNNL p. 273 Eq 7.44

$$K_c := \frac{1}{\sqrt{n-1}} \cdot K_s$$

< K_c is a cleverly scaled version of K_s

$$S_{Kc} := \frac{1}{n-1} \cdot K_c^T \cdot \left(I - \frac{1}{n} \cdot J \cdot J^T \right) \cdot K_c \qquad S_{Kc} = \begin{pmatrix} 0.05 & 0.0391 & 0.0472 \\ 0.0391 & 0.05 & 0.0418 \\ 0.0472 & 0.0418 & 0.05 \end{pmatrix}$$

^ Variance-Covariance matrix of Kc is generally meaningless...

$$K_c^T \cdot K_c = \begin{pmatrix} 1 & 0.7813 & 0.9446 \\ 0.7813 & 1 & 0.8358 \\ 0.9446 & 0.8358 & 1 \end{pmatrix}$$

< However, Sum of Squares of K_c produces the Correlation Matrix! The calculation is interesting, but I'm uncertain whether this result is all that important.

Prototype in R:

```
COMMANDS:
#SCALE FUNCTION PROTOTYPE
#AND cov() cor() FUNCTIONS
#MEAN CENTER DATA:
Km=scale(K,center=T,scale=F)
Km
#STANDARDIZE DATA:
Ks=scale(K,center=T,scale=T)
Ks
#VARIANCE-COVARIANCE MATRIX:
S=cov(K)
S
#CORRELATION MATRIX:
R=cor(K)
R
```

< scale() allows one to mean-center and standardize data among other things...

< cov() produces Variance-Covariance matrix

< cor() produces Correlation matrix

^ all values verified but not shown here.