

Demonstration of the Importance of Centering Continuous Variables Before Creating Interaction Terms

Scot W. McNary

June 20, 2003

Aiken & West [1] emphasize it is important to center continuous variables before creating a product term with them to represent an interaction. This applies to software in which you must create your own interaction terms before running a model (e.g., SAS Proc Reg) or software in which the interaction terms are created for you by specifying the interaction component in the model statement (e.g., SAS Proc Mixed: $y = x1 \ x2 \ x1 * x2$). This is a demonstration why. First, let's create some variables (there are a couple of options to creating low correlations between variables):

```
> x1 <- 10 + rnorm(100)
> x2 <- rnorm(length(x1), mean = mean(x1), sd = sd(x1))
> x2 <- rnorm(length(x1), mean = 10 + 1e-05 * x1, sd = sd(x1))
```

Check their correlation:

```
> cor(x1, x2)
[1] 0.1725266
```

Create the product term and combine variables:

```
> x1.x2 <- x1 * x2
> unc <- data.frame(x1, x2, x1.x2)
> apply(unc, 2, mean)
```

x1	x2	x1.x2
9.969982	10.028485	100.121214

```
> apply(unc, 2, sd)
```

x1	x2	x1.x2
0.9242540	0.8703716	13.7975316

Look at the SD of the product term.

Now for the intercorrelations:

```
> cor(unc)
```

```

          x1          x2          x1.x2
x1      1.0000000 0.1725266 0.7820883
x2      0.1725266 1.0000000 0.7462249
x1.x2  0.7820883 0.7462249 1.0000000

```

The size of the correlations between the main effects and the product term are very large relative to the small correlation between the main effects.

Now mean center them and add the centered variables to the original data set:

```

> cent <- data.frame(x1c = x1 - mean(x1), x2c = x2 - mean(x2))
> cent$x1c.x2c <- cent$x1c * cent$x2c
> names(cent)

[1] "x1c"      "x2c"      "x1c.x2c"

```

The means have changed and the SD of the product term has decreased in size, but the SD for x1 and x2 remain the same:

```

> apply(cent, 2, mean)

          x1c          x2c          x1c.x2c
-2.522427e-15 -3.801404e-15  1.374002e-01

> apply(cent, 2, sd)

          x1c          x2c          x1c.x2c
0.9242540 0.8703716 0.7802829

```

Now, re-check the intercorrelations:

```

> cor(cent)

          x1c          x2c          x1c.x2c
x1c      1.00000000 0.17252663 0.03191844
x2c      0.17252663 1.00000000 0.02479449
x1c.x2c  0.03191844 0.02479449 1.00000000

```

Note the correlations between main effects and product terms are much smaller with centered versions.

Check on multicollinearity with usual diagnostics:

```

> y <- 10 + rnorm(100)
> library(Design)
> summary(lm(y ~ x1 + x2 + x1.x2, data = unc))

```

```

Call:
lm(formula = y ~ x1 + x2 + x1.x2, data = unc)

```

```

Residuals:
      Min       1Q   Median       3Q      Max

```

-2.81361 -0.78282 -0.04889 0.71682 2.04575

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	16.17054	13.16636	1.228	0.222
x1	-0.59300	1.31637	-0.450	0.653
x2	-0.74849	1.30856	-0.572	0.569
x1.x2	0.07212	0.13048	0.553	0.582

Residual standard error: 1.012 on 96 degrees of freedom
Multiple R-Squared: 0.01742, Adjusted R-squared: -0.01329
F-statistic: 0.5673 on 3 and 96 DF, p-value: 0.6379

Note the SEs from this model.

Check vifs from this model (over 30 are problematic):

```
> vif(lm(y ~ x1 + x2 + x1.x2, data = unc))
```

	x1	x2	x1.x2
	139.4872	122.2568	216.3384

In contrast, look at the same model using centered effects:

```
> summary(lm(y ~ x1c + x2c + x1c.x2c, data = cent))
```

Call:

```
lm(formula = y ~ x1c + x2c + x1c.x2c, data = cent)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.81361	-0.78282	-0.04889	0.71682	2.04575

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	9.96307	0.10280	96.915	<2e-16 ***
x1c	0.13026	0.11179	1.165	0.247
x2c	-0.02945	0.11869	-0.248	0.805
x1c.x2c	0.07212	0.13048	0.553	0.582

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.012 on 96 degrees of freedom
Multiple R-Squared: 0.01742, Adjusted R-squared: -0.01329
F-statistic: 0.5673 on 3 and 96 DF, p-value: 0.6379

Notice the difference in SEs for the main effect terms between centered and uncentered models. Now check vifs for the model with uncentered effects:

```
> vif(lm(y ~ x1c + x2c + x1c.x2c, data = cent))
```

	x1c	x2c	x1c.x2c
	1.031467	1.031062	1.001361

Much smaller!

Look a little more directly at the variances for the terms in the two models:

```
> vc <- diag(summary(lm(y ~ x1c + x2c + x1c.x2c, data = cent))$cov.unscaled)
> vu <- diag(summary(lm(y ~ x1 + x2 + x1.x2, data = unc))$cov.unscaled)
```

Indeed, the variances from the uncentered model are huge!

```
> vu/vc
```

(Intercept)	x1	x2	x1.x2
16403.2054	138.6484	121.5489	1.0000

References

- [1] Aiken, L. West, S.(1991) *Testing Interactions in Multiple Regression*, Lawrence Erlbaum.