This example illustrates modeling an interaction with centering and transformations. Our goal is to build a model for psychology graduate faculty salary as a function of years in tenure track and program level (MA only or PhD) of academic employment.

These are real data from the APA Research Office, taken from a 2005 study of salaries of faculty in graduate psychology programs as described in MRC08. The data set is available at http://wfs.cgu.edu/bergerd/Data/APAsal05.sav .

In the earlier analysis we learned that salary is related to the number of years faculty members are in a tenure track position (**yearstt**) and whether they teach at an MA only or PhD institution (**level**), and we found a statistically significant interaction between these two variables such that the relationship between **yearstt** and **salary** is greater at PhD institutions than at MA only institutions. After we control for time on the job, **sex** is not predictive of **salary** and there are no interactions with **sex**. Consequently, we will not include **sex** in the model for predicting **salary**.

We also found that compared to **salary**, the log of salary (**lnsal**) showed a stronger relationshp with **yearstt** and there were fewer outliers. For this reason we will model **lnsal**, but then convert back to **salary** to present our findings to a lay audience.

To measure program level, we will code **leveld** as MA = 0 and PhD = 1. We will limit the analysis to faculty in tenure track positions. For illustration we will first analyze the data without centering and then with centering to allow a comparison of the results.

The first step is to create a new variable, **levxyear**, to serve as an interaction term. We do this by multiplying **leveld** by **yearstt**. In the regression analysis it is important to enter the main effects prior to the interaction term so that we can test the contribution of the interaction beyond those main effects. Below is the syntax; on the next page are the point and click instructions.

```
use all.
split file off.
*compute term for interaction between leveld and yearstt.
compute levxyear = leveld*yearstt.

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA COLLIN TOL CHANGE ZPP
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT lnsal
  /METHOD=ENTER yearstt
  /METHOD=ENTER leveld
  /METHOD=ENTER levxyear
  /SCATTERPLOT=(*ZRESID ,*ZPRED)
  /RESIDUALS HIST(ZRESID) NORM(ZRESID)
  /CASEWISE PLOT(ZRESID) OUTLIERS(3).
```

To assure that any earlier split file is turned off.

To generate the syntax on the previous page, click Transform, Compute Variable…, in the Target Variable: window enter **levxyear** and in the Numeric Expression: window enter **leveld*yearstt**, click OK. For the regression analysis, click Analyze, Regression, Linear…, then select **lnsal** as the Dependent variable, select **yearstt** as the first Independent, click Next, select **leveld** as the next Independent variable, click Next, and select **levxyear** as the last Independent variable. Click Statistics… and select Estimates, Model fit, R squared change, Descriptives, Part and partial correlations, Collinearity diagnostics, and Casewise diagnostics with Outliers 3 standard deviations. Click Continue, Plots…, select ***ZRESID** as the Y variable and ***ZPRED** as the X variable, check Histogram and Normal probability plot, click Continue, and OK to run (or Paste to save the syntax).

Let's check the output file:

Descriptive Statistics

	Mean	Std. Deviation	N
= log of salary	11.1581	.31233	4558
Estimated years in tenure track	13.1709	8.26314	4558
leveld (MA=0;PhD=1)	.83	.376	4558
levxyear	11.1129	9.05041	4558

Here we see that we have all of the cases we expect and that the means are as we expect. No apparent problems here. We checked the shapes of the distributions earlier, which showed a slight skew in **lnsal**, but it doesn't look very serious.

Correlations

		= log of salary	Estimated years in tenure track	leveld (MA=0; PhD=1)	levxyear
Pearson Correlation	= log of salary	1.000	.745	.267	.719
	Estimated years in tenure track	.745	1.000	.059	.789
	leveld (MA=0;PhD=1)	.267	.059	1.000	.556
	levxyear	.719	.789	.556	1.000
Sig. (1-tailed)	= log of salary	.	.000	.000	.000
	Estimated years in tenure track	.000	.	.000	.000
	leveld (MA=0;PhD=1)	.000	.000	.	.000
	levxyear	.000	.000	.000	.
N	= log of salary	4558	4558	4558	4558
	Estimated years in tenure track	4558	4558	4558	4558
	leveld (MA=0;PhD=1)	4558	4558	4558	4558
	levxyear	4558	4558	4558	4558

Here we see the strong correlation of **lnsal** with **yearstt** (.745) and the somewhat weaker correlation with **leveld**. The correlation of **lnsal** with the interaction term **levxyear** is not very useful when the interaction is constructed with raw (not centered) predictor variables.

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.745[a]	.554	.554	.20852	.554	5.667E3	1	4556	.000
2	.777[b]	.604	.604	.19657	.050	571.680	1	4555	.000
3	.779[c]	.607	.606	.19594	.003	30.499	1	4554	.000

a. Predictors: (Constant), Estimated years in tenure track

b. Predictors: (Constant), Estimated years in tenure track, leveld (MA=0;PhD=1)

c. Predictors: (Constant), Estimated years in tenure track, leveld (MA=0;PhD=1), levxyear

d. Dependent Variable: = log of salary

This hierarchical analysis confirms that the interaction term does contribute significantly beyond the main effects, but we note that **yearstt** is the strongest predictor by far, accounting for 55.4% of the variance in **lnsal**, while **leveld** contributes an additional 5.0%. The interaction term **levxyear** contributes only .003 or 0.3%. While this is statistically significant, it may not be practically important. Nonetheless, we will include it in our model.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	10.787	.006		1855.992	.000					
	Estimated years in tenure track	.028	.000	.745	75.282	.000	.745	.745	.745	1.000	1.000
2	(Constant)	10.640	.008		1290.225	.000					
	Estimated years in tenure track	.028	.000	.731	78.298	.000	.745	.757	.730	.996	1.004
	leveld (MA=0; PhD=1)	.186	.008	.223	23.910	.000	.267	.334	.223	.996	1.004
3	(Constant)	10.692	.013		853.944	.000					
	Estimated years in tenure track	.023	.001	.617	27.222	.000	.745	.374	.253	.168	5.951
	leveld (MA=0; PhD=1)	.122	.014	.146	8.743	.000	.267	.128	.081	.308	3.248
	levxyear	.005	.001	.150	5.523	.000	.719	.082	.051	.116	8.587

a. Dependent Variable: = log of salary

Model 3 shows the coefficients for a regression model to predict **lnsal**. First, we note that both **leveld** and **levxyear** are zero for MA faculty, so Model 3 for MA faculty simplifies to 10.692 + .023 * (**yearstt**). For PhD faculty, **leveld**=1, so the B weight of .122 is added to the prediction of **lnsal** for all PhD faculty. Because **leveld**=1 for PhD faculty, **levxyear** is equal to **yearstt** for PhD faculty. Thus, the total weight given to **yearstt** for PhD faculty is .023 +.005. The model for PhD faculty alone is (10.692 + .122) + (.023+.005)*__yearstt__, which is 10.814 + .028*__yearstt__. Thus, the B weight for the interaction term (.005) is the difference in the weight given to **yearstt** for MA faculty (weight = .023) compared to PhD faculty (weight = .028). The significance test of B for **levxyear** is a test of whether the increment .005 is statistically significantly greater than zero. Average salary is greater for PhD faculty than for MA faculty with the same number of year in tenure track, and years in tenure track is more strongly related to salary for PhD faculty than for MA faculty. (You may need to read this paragraph again! Practice explaining it in your own words. A graph could be very useful, as we will soon see.)

Because the interaction term is the product of **yearstt** and **leveld** (and both variables have no negative values), the product **levxyear** is strongly correlated with each of these variables. Hence, the tolerance for all three variables is low when the interaction term is included on the third step. The tests of significance for these main effects in Model 3 are not interpretable because of the overlap with the interaction term. We may be interested in the contribution of the interaction beyond the main effects, but main effects beyond the interaction is less interpretable.

We asked SPSS to give us casewise information on outliers that are 3 or more standard errors from the predicted value. If the errors were normally distributed, what proportion of the standardized residuals would we expect to be greater than 3 standard errors? From a Z table or from StatWISE we find this proportion to be .00135 on each tail, or .00270 for the two tails. How many cases would we expect to be 3 or more standard errors from prediction? This analysis included 4558 cases, and (.00270 * 4558) =12.3, so we expect about a dozen. Below is the first portion of a list of cases with Std. Residual greater than 3 in absolute value (+ or – 3).

Casewise Diagnostics[a]

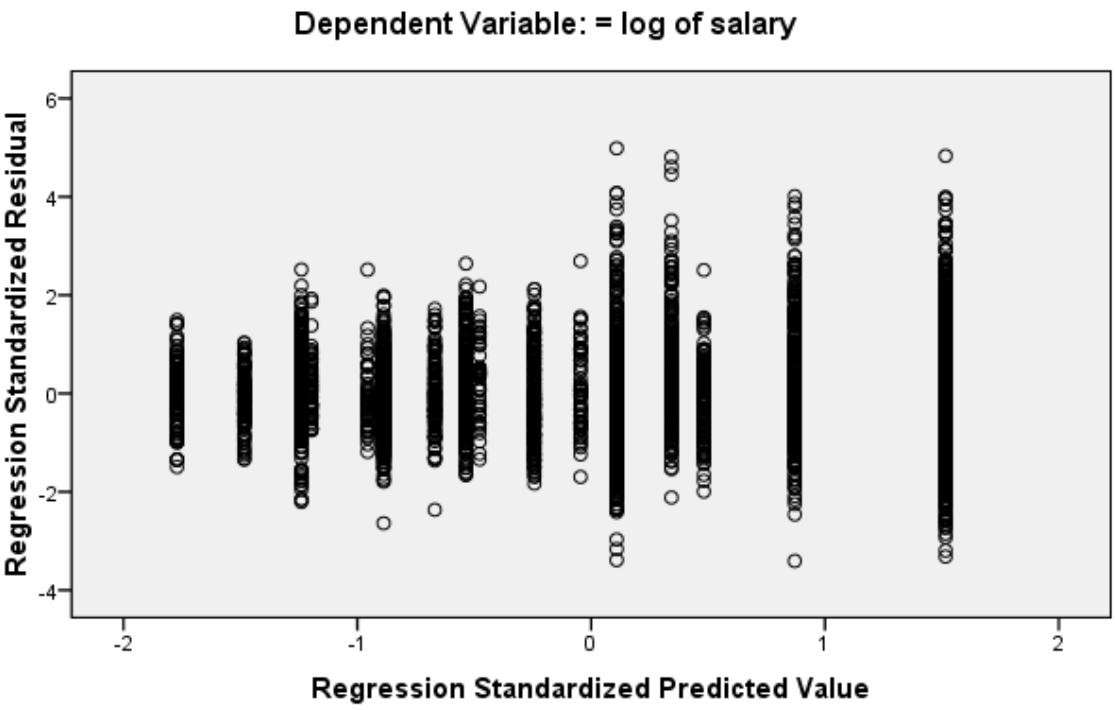
Case Number	Std. Residual	= log of salary	Predicted Value	Residual
282	3.953	12.30	11.5268	.77457
288	3.396	11.85	11.1846	.66536
300	3.467	12.21	11.5268	.67926
417	3.240	12.16	11.5268	.63479
427	3.282	12.17	11.5268	.64312

The entire list has 46 such cases, where the maximum standardized residual is 4.986. The implication is that the residuals are not normally distributed. We also note that only 5 of the 46 outliers are negative, so we know we have positive skew. Because our sample is very large (N=4558), the departure from normality may not have a great impact, but it would be prudent to do some sensitivity analyses to see how much our models would be affected if we trimmed a portion of the outliers or Winsorized by setting the k most extreme cases on each end equal to the value of the case k steps from the end. For example, we might set the nine largest values equal to the tenth largest value and the nine smallest values equal to the tenth smallest value. Comparison of the original analysis with an analysis of the Winsorized data allows us to assess the impact of the most extreme cases on the original analysis.

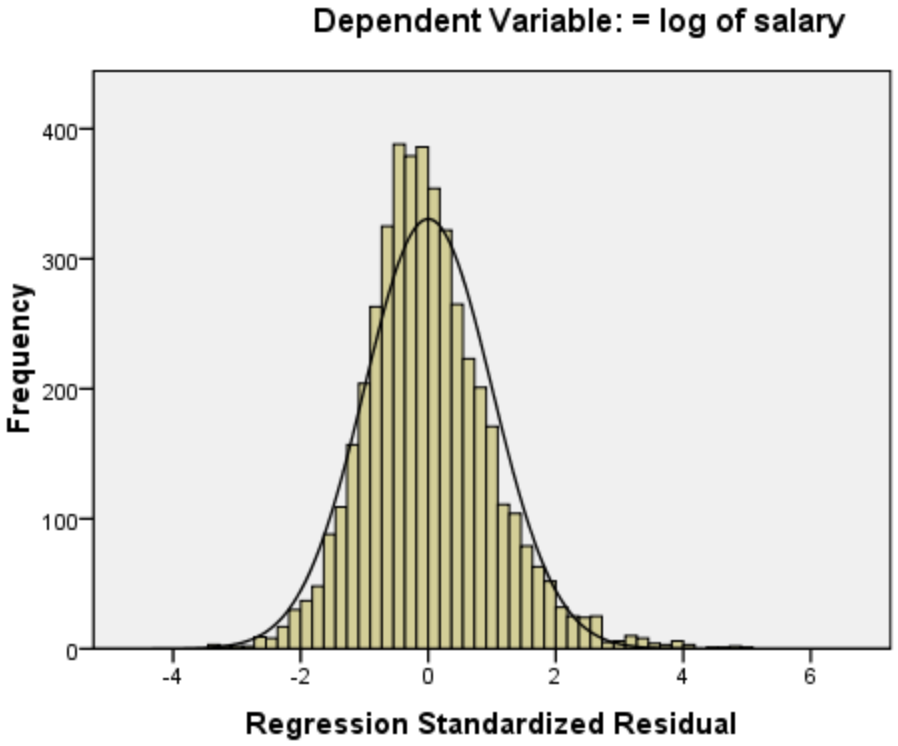
It is useful to examine the scatterplot of standardized residuals as a function of standardized predicted values (shown on the next page). We see several distinct vertical plots, 16 of them to be exact. Why do we have only 16 X values? What do they represent?

We are using two predictor variables. The first, **leveld,** has two categories (MA, PhD) while the second, **yearstt,** has eight categories. The combination produces 16 distinct combinations of these two predictor values. The rightmost vertical plot represents Full Professors who have the most years on the job at PhD granting institutions, and the leftmost plot represents beginning faculty at MA granting institutions.

We can draw a horizontal line through the zero value to help assess linearity and homoscedasticity. The plots look reasonably normal, or at least there are no extreme outliers. There appears to be somewhat greater variance for cases with greater predicted values. Categories with more cases tend to show greater spread, so we should take that into consideration as we interpret the spread. We do have more cases in the two top groups.



The histogram of standardized residuals combines residuals from all levels of the predicted value. It may help to imagine that the plot above is rotated 90 degrees clockwise and all points drop to the bottom, creating a distribution centered on zero and ranging from -4 to 6.



This plot shows an interesting small clump of residuals on the right tail. Perhaps these are faculty with supplemented salaries, maybe from administrative work or grant work? Maybe Stanford?

Centered vs. uncentered predictors

A centered variable is created by subtracting the mean of the original variable from every original observation. The centered variable has a mean of zero, with scores indicating deviations from the mean, positive or negative. There are advantages for centering predictors, especially when we wish to test interactions in a regression analysis and interpret B and beta values.

When we create an interaction term by multiplying together two uncentered main effects, we are likely to observe a large correlation between the interaction term and the two main effects. When we enter the main effects and interaction term into a model together, these variables may have a high degree of multicollinearity with each other. Note in the Coefficients table how much the tolerance for the main effects decreases from Model 2 to Model 3, when the interaction term is added to the model. Tolerance is (one minus multicollinearity).

In general, the regression weights and the tests of statistical significance for uncentered main effects are not easily interpreted when the interaction term is included in the model. The null hypothesis for each t-test of B (or beta) weights is that the variable contributes to the prediction of Y beyond the contribution of all other variables in the model. In Model 3, the test of a main effect is the test of that main effect beyond the other main effect and also beyond the interaction. Ordinarily it makes much more sense to interpret the contribution of the interaction beyond the main effects than to interpret main effects beyond the interaction with uncentered predictors. The t-tests of statistical significance for the main effects are likely to be greatly reduced in the presence of the interaction because the interaction overlaps so much with main effects.

The intercept can be interpreted as the predicted value of the criterion variable when all predictors are zero. With uncentered data, zero may not be interesting or even meaningful. In our example, the intercept in Model 3 in the Coefficients table is 10.692. That is the predicted value of **lnsal** for someone in a Master's institution (**leveld** = 0) who has zero years in a tenure track position (**yearstt** = 0). However, the minimum value for **yearstt** in our data set is 1.5, for people in their first or second year. No one has zero years. This problem would be more salient if we used a predictor like GRE score, because then the intercept would be the predicted value for someone with a score of zero on GRE. That is just silly.

To minimize rounding error when we center, we should use the mean computed to many places of accuracy. This is easily done in SPSS. Go to the Descriptive Statistics near the beginning of the regression output. The mean for **leveld** is shown as .83 (verify N=4558). To display the mean with more places of accuracy, double-click on the table, and then double-click on the mean, to see 0.8297498903027644. If you hold Ctrl while you press c (i.e., Ctrl-c) while the number is highlighted, the number will be stored in your temporary clipboard. You can paste this value into your SPSS syntax by holding Ctrl while pressing v (i.e., Ctrl-v). For example, we can create the centered variable for level by entering the following command into the syntax window:

compute leveldc = leveld - 0.8297498903027644. (I pressed Ctrl-v to insert the mean.)

Alternatively, click Transform, Compute Variable…, in the Target Variable: window enter **leveldc** and in the Numeric Expression: window enter **leveld - 0.8297498903027644**, click OK. Rather than entering the 16 digits one-by-one, you can use the Ctrl-v trick here, too.

Similarly, center yearstt: **compute yearsttc = yearstt - 13.170908293111014.**
compute ycxlc = yearsttc * leveldc.

It is instructive to compare the correlations for centered and uncentered variables.

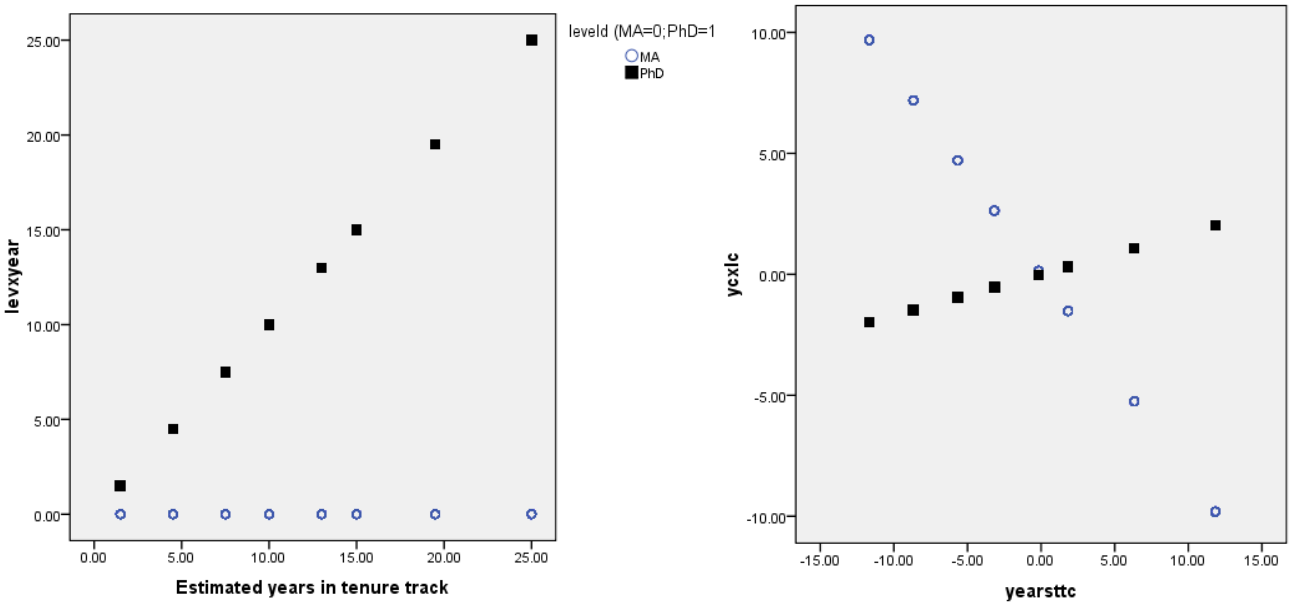
Correlations[a]

		= log of salary	Estimated years in tenure track	yearsttc	leveld (MA=0; PhD=1)	leveldc	levxyear	ycxlc
= log of salary	Pearson Correlation	1	.745**	.745**	.267**	.267**	.719**	.026
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.075
Estimated years in tenure track	Pearson Correlation	.745**	1	1.000**	.059**	.059**	.789**	-.002
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.898
yearsttc	Pearson Correlation	.745**	1.000**	1	.059**	.059**	.789**	-.002
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.898
leveld (MA=0; PhD=1)	Pearson Correlation	.267**	.059**	.059**	1	1.000**	.556**	-.104**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
leveldc	Pearson Correlation	.267**	.059**	.059**	1.000**	1	.556**	-.104**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000
levxyear	Pearson Correlation	.719**	.789**	.789**	.556**	.556**	1	.285**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000
ycxlc	Pearson Correlation	.026	-.002	-.002	-.104**	-.104**	.285**	1
	Sig. (2-tailed)	.075	.898	.898	.000	.000	.000	

**. Correlation is significant at the 0.01 level (2-tailed).

a. Listwise N=4558

Notice that the correlations between level and years are the same with the uncentered and centered main effects (top dashed oval), but the correlations with the interaction terms computed from centered vs. uncentered main effects differ markedly (bottom dashed oval). The correlation with main effects is much larger for interaction terms based on the uncentered variables than for interactions based on centered variables (e.g., the correlations with **yearstt** are .789 vs. -.002)



As always, a figure is worth a thousand words. The figures show why the correlation with **yearstt** is smaller for the centered interaction term than for the interaction term based on uncentered predictors. (Recall that there are more cases in the PhD condition – the dark squares.)

Now we are ready to run the regression analysis with the centered variables. Replace **yearstt** with **yearsttc**, replace **leveld** with **leveldc**, and replace **levxyear** with **ycxlc** in the hierarchical analysis where we enter one variable on each step.

Descriptive Statistics

	Mean	Std. Deviation	N
= log of salary	11.1581	.31233	4558
yearsttc	.0000	8.26314	4558
leveldc	.0000	.37589	4558
ycxlc	.1843	3.10542	4558

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.745[a]	.554	.554	.20852	.554	5.667E3	1	4556	.000
2	.777[b]	.604	.604	.19657	.050	571.680	1	4555	.000
3	.779[c]	.607	.606	.19594	.003	30.499	1	4554	.000

a. Predictors: (Constant), yearsttc

b. Predictors: (Constant), yearsttc, leveldc

c. Predictors: (Constant), yearsttc, leveldc, ycxlc

d. Dependent Variable: = log of salary

Compare this model summary to the earlier model summary, and you will see that they are essentially identical. That is, the contribution of each term in the hierarchical model is the same whether we use uncentered or centered predictors. The difference shows up in the Coefficients table, especially in the B coefficients, correlations with the interaction term, and tolerance.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	11.158	.003		3612.622	.000					
	yearsttc	.028	.000	.745	75.282	.000	.745	.745	.745	1.000	1.000
2	(Constant)	11.158	.003		3832.205	.000					
	yearsttc	.028	.000	.731	78.298	.000	.745	.757	.730	.996	1.004
	leveldc	.186	.008	.223	23.910	.000	.267	.334	.223	.996	1.004
3	(Constant)	11.157	.003		3837.435	.000					
	yearsttc	.028	.000	.731	78.527	.000	.745	.758	.730	.996	1.004
	leveldc	.190	.008	.229	24.432	.000	.267	.340	.227	.986	1.015
	ycxlc	.005	.001	.052	5.523	.000	.026	.082	.051	.989	1.011

a. Dependent Variable: = log of salary

Consider Model 3. The B for the constant is 11.157. This can be interpreted as the predicted value of **lnsal** when all predictors are equal to zero. With centered variables, this means someone with average number of years in tenure track and an average person across the two values of **leveld**. Thus, the constant is the average value of **lnsal** for the sample (note that the mean in Descriptives = 11.158). If we wish to use the Unstandardized Coefficients for centered variables to predict **lnsal** for a specific case, we need to convert **yearstt** to **yearsttc** by subtracting the mean of **yearstt**, and convert **leveld** to **leveldc** by subtracting the mean for **leveld**.

The beta weights in the final model are more interpretable with centered variables than with uncentered predictors. Yet, when we have a variable like **leveld** where 0 is meaningful, we may choose not to center it because the values of 0 and 1 are easier to use than the centered values.

Graphing interactions

In this example, the interaction is statistically significant. Is it large enough to be practically significant? The answer to that question will depend on our goals. If we are interested for theoretical reasons in whether an interaction exists, then even a very small effect may be of theoretical interest. In our example we are building a model for prediction of salaries. Here a small effect may look unimportant; yet the presence of an interaction may have greater consequences for predictions at the extremes of the range, where predictions can diverge more.

In addition to statistical significance, it is always important to describe the size and direction of an interaction. This can be done mathematically with the table of coefficients, but often it is desirable to present interactions with a figure. Interactions can be plotted using centered variables or uncentered variables, but it is likely to be easier to interpret figures with uncentered variables. In our example, **leveld** takes on values of 0 for MA institutions and 1 for PhD institutions, while the centered version **leveldc** has the value of -0.8297498903027644 for MA institutions and + 0.17025010969723564 for PhD institutions, values that are not easy to use.

Excel can produce useful figures using copy-and-paste information from SPSS analyses. Here we will use the Excel file **Plotting Regression Interactions.xls,** which is available in our class website, to help construct a figure. In this example we have only two predictors (uncentered **yearstt** and **leveld**) and their interaction (**levxyear**) for the dependent variable **lnsal**. To make our figure more useful, we will convert the predicted **lnsal** values back to **salary** values. A worksheet to help us is on the *Log to raw* tab.

On the *Log to raw* tab we find labels (in orange cells) for the constant and the three predictor variables in the order they appear in the SPSS output, followed by the corresponding B weights in the yellow cells. The blue cells indicate values for the predictor variables that we wish to plot. In the example, I chose to plot values for 1, 15, and 30 years in tenure track for each of the two institution levels. The worksheet computes the predicted value of lnsal and converts that value into salary (by raising e to the value of lnsal).

		B weight	AB11	AB21	AB31	AB21	AB22	AB32
	constant	10.692	1	1	1	1	1	1
A	yearstt	.023	1	15	30	1	15	30
B	level	.122	0	0	0	1	1	1
AB	levxyear	.005	0	0	0	1	15	30
	predicted	lnsal	10.71555	11.04212	11.39202	10.84241	11.24164	11.66939
		salary	45050.93	62449.93	88611.63	51144.47	76240.12	116937.3

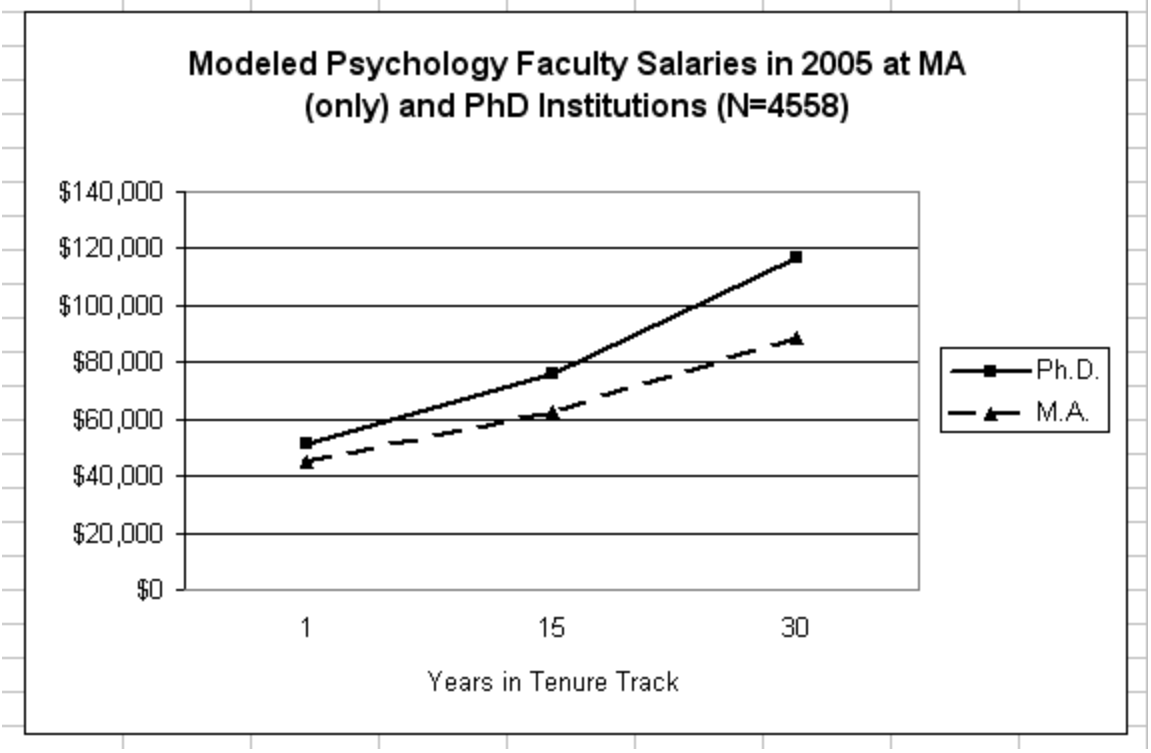
The B weight coefficient can be copied from SPSS with copy-and-paste if you have the SPSS output file open. With the Excel *Log to raw* tab open, go to the SPSS output. You can toggle to active files by holding **Alt** and pressing **Tab**. Go to Model 3 in the SPSS output file Coefficients table to locate the desired B weights. Double-click on the table to open Pivot Table Coefficients and then left-click on the B weight for the constant to highlight the number. Now hold the **Shift** key and left-click on the last coefficient to highlight all of the coefficients for Model 3. Now hold the **Ctrl** key and press **c** to copy the highlighted information to your clipboard. Next go to the Excel *Log to raw* page, left-click on the cell under the B weight column next to the constant, hold **Ctrl** and press **v** to insert the numbers from SPSS. Although the numbers may appear with

only a few decimal places, you will see that about 15 significant digits have been copied from SPSS if you left-click on a number.

The modeled salary values are copied into a 2x3 table for graphing, and the values appear in the graph below.

Here we notice that for those in a tenure track for 1 year, the difference between salaries at M.A. and Ph.D. institutions is 51144 – 45051 = $6,093, but for those in tenure track for 30 years the difference is $28,325. Small coefficients in regression can have a large effect if they act on a variable with large variance. Here the change in scale from log to raw salaries also contributes to the larger difference. Note that the model shows only major patterns, not details in the data.

	Years in Tenure Track		
	1	15	30
M.A.	45051	62450	88612
Ph.D.	51144	76240	116937



Variations: With categorical variables, it may be better to use bar graphs. With two continuous variables, you may choose to plot the relationship between one predictor and the criterion for each of three levels of the other predictor. Those three levels might be chosen as the mean and values one standard deviation above and below the mean, or you may choose other values that are easier to interpret. If you have used a log transformation you should plot more than two values on the horizontal axis to show the curvature in the model. Excel is very flexible, allowing you to edit figures in many ways.

BIG CAUTION! Check everything to make sure it is working correctly. With Excel it is terribly easy to have an error that is hidden in the code. It is easy to write over a formula by mistake or to enter a number in the wrong place or ….