Expanding the Model Capabilities: Dummy Variables, Interactions, and Nonlinear Transformations (Soc 504)

Until now, although we have considered including multiple independent variables in our regression models, we have only considered continuous regressors and linear effects of them on the outcome variable. However, the linear regression model is quite flexible, and it is fairly straightforward to include qualitative variables, like race, gender, type of occupation, etc. Furthermore, it is quite easy to model nonlinear relationships between independent and dependent variables.

1 Dummy Variables

When we have dichotomous (binary) regressors, we construct something called 'indicator,' or 'dummy' variables. For example, for gender, we could construct a binary variable called 'male' and code it 1 if the person is male and 0 if the person is female. Suppose gender were the only variable in the model:

$$\hat{Y}_i = b_0 + b_1 Male_i.$$

The expected value of Y for a male, then, would be $b_0 + b_1$, and the expected value of Y for a female would be just b_0. Why? Because $0 \times b_1 = 0$, and $1 \times b_1$ is just b_1. Interestingly, the t-test on the b_1 parameter in this case would be identical to the t test you could perform on the difference in the mean of Y for males and females.

When we have a qualitative variable with more than two categories (e.g., race coded as white, black, other), we can construct more than one dummy variable to represent the original variable. In general, the rule is that you construct $k-1$ dummy variables for a qualitative variable with k categories. Why not construct k dummies? Let's assume we have a model in which race is our only predictor. If we construct a dummy for 'black' and a dummy for 'other,' then we have the following model:

$$\hat{Y}_i = b_0 + b_1 Black_i + b_2 Other_i.$$

When 'black'=1, we get:

$$\hat{Y}_i = b_0 + b_1.$$

When 'other'=1, we get:

$$\hat{Y}_i = b_0 + b_2.$$

But, when we come across an individual who is white, both dummy coefficients drop from the model, and we're left with:

$$\hat{Y}_i = b_0.$$

If we also included a dummy variable for 'white,' we could not separate the effect of the dummy variable and the intercept (more specifically, there is perfect negative collinearity between them). We call the omitted dummy variable the 'reference' category, because the other dummy coefficients are interpreted in terms of the expected mean change relative to it. (note: as an alternative to having a references group, we could simply drop the intercept from the model.)

When we have only dummy variables in a model representing a single qualitative variable, the model is equivalent to a one-way ANOVA. Recall from your previous coursework that ANOVA is used to detect differences in means across two or more groups. When there are only two groups, a simple t-test is appropriate for detecting differences in means (indeed, an ANOVA model—and a simple regression as discussed above—would yield an F value that would simply be t^2). However, when there are more than two groups, the t-test becomes inefficient. We could, for example, conduct a t-test on all the possible two-group comparisons, but this would be tedious, because the number of these tests is $\left(\begin{array}{c} k \\ 2 \end{array} \right) = \frac{k!}{(k-2)!2!}$, where k is the number of categories/groups. Thus, in those cases, we typically conduct an ANOVA, in which all group means are simultaneously compared. If the F statistic from the ANOVA is significant, then we can conclude that at least one group mean differs from the others.

The standard ANOVA model with J groups is constructed using the following computations:

$$\begin{aligned}
Grand\ Mean \equiv \bar{\bar{X}} &= \frac{\sum_j \sum_i X_{ij}}{n} \\
Sum\ Squares\ Total\ (SST) &= \sum_j \sum_i \left(X_{ij} - \bar{\bar{X}}\right)^2 \\
Sum\ Squares\ Between\ (SSB) &= \sum_j n_j \left(\bar{X}_j - \bar{\bar{X}}\right)^2 \\
Sum\ Squares\ Within\ (SSW) &= \sum_j \sum_i \left(X_{ij} - \bar{\bar{X}}_j\right)^2
\end{aligned}$$

The ANOVA table is then constructed as:

SS	df	MS	F
Between	J-1	$\frac{SSB}{df}$	F
Within	N-J	$\frac{SSW}{df}$	
Total	N-1		

This should look familiar—it is the same table format used in regression. Indeed, as I've said, the results will be equivalent also. The TSS calculation is identical to that from regression. The degrees of freedom are also equivalent. Furthermore, if we realize that in a model with dummy variables only, then $\hat{Y}$ is simply the group mean, then the SSW

calculation is identical to the $(Y - \hat{Y})$ calculation we use in regression. Thus, ANOVA and regression are equivalent models, and when dummy variables are used, the regression model is sometimes called ANCOVA (analysis of covariance).

A limitation of the ANOVA approach is that the F-test only tells us whether at least one group mean differs from the others; it doesn't pinpoint *which* group mean is the culprit. In fact, if you use ANOVA and want to determine which group mean differs from the others, additional tests must be constructed. However, that's the important feature of regression with dummies—the t-tests on the individual dummy coefficients provides us with this information. So, why do psychologists (and clinicians, often) use ANOVA? A short answer is that there is a 'cultural inertia' at work. Psychologists (and clinicians) historically have dealt with distinct groups in clinical trials/experimental settings in which the ANOVA model is perfectly appropriate, and regression theoretically unnecessary. Sociologists, on the other hand, don't use such data, and hence, sociologists have gravitated toward regression analyses.

So far, we have discussed the inclusion of dummy variables in a relatively simple model. When we have continuous covariates also in the model, interpretation of the dummy variable effect is enhanced with graphics. A dummy coefficient in such a model simply tells us how much the regression line 'jumps' for one group versus another. For example, suppose we were interested in how education relates to depressive symptoms, net of gender. Then, we might construct a model like:

$$\hat{Y} = b_0 + b_1 Education + b_2 Male.$$

I have conducted this model using data from the National Survey of Families and Households (NSFH), with the following results: $b_0 = 25.64$, $b_1 = -.74$, and $b_2 = -3.27$. These results suggest that education reduces symptoms (or more correctly, that individuals with greater education in the sample have fewer symptoms—at a rate of $-.74$ depressive symptoms per year of education) and that men have 3.27 fewer symptoms on average. Graphically, this implies:

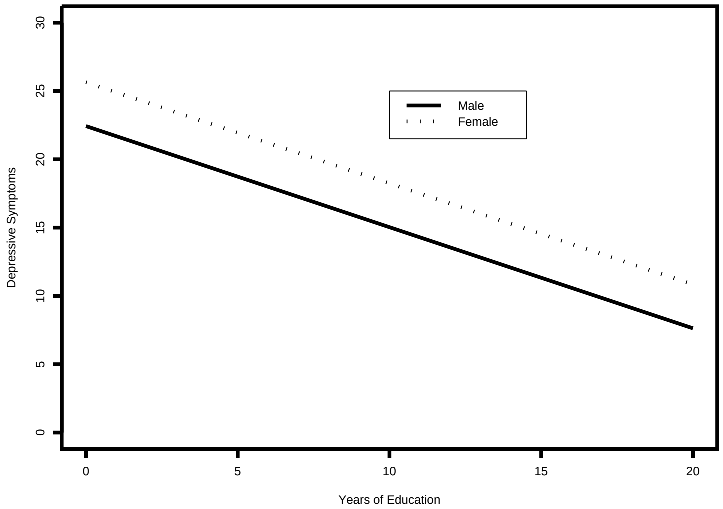

Figure 1. Depressive Symptoms by Years of Education and Gender

If we had a qualitative variable in the model that had more than two categories (hence more than one dummy variable), we would have more than two lines, but all would be parallel.

2 Statistical Interactions

Notice that the lines in the above example are parallel—that it is assumed in this model that education has the same effect on depressive symptoms for men and for women. This is often an unrealistic assumption, and often our theories provide us with reasons why we may expect different slopes for different groups. In these cases, we need to come up with some way of capturing this difference in slopes.

2.1 Interactions Between Dummy and Continuous Variables

The simplest interactions involve the interaction between a dummy variable and a continuous variable. In the education, gender, and depressive symptoms example, we may expect that education's effect varies across gender, and we may wish to model this differential effect. A model which does so might look like:

$$\hat{Y} = b_0 + b_1 Education + b_2 Male + b_3(Education \times Male)$$

This model differs from the previous one, because it includes a 'statistical interaction term,' $Education \times Male$. This additional variable is easy to construct—it is simply the multiple of each individual's education value and their gender. For women ($Male = 0$), this interaction term is 0, but for men, the interaction term is equal to their education level ($Education \times 1$). This yields the following equations for men and women:

$$\begin{aligned} \hat{Y}_{men} &= (b_0 + b_2) + (b1 + b_3)Education \\ \hat{Y}_{women} &= b_0 + b_1 Education \end{aligned}$$

In the equation for men, I have consolidated the effects of education into one coefficient, so that the difference in education slopes for men and women is apparent. We still have an expected mean difference for men and women (b_2), but we now also allow the effect of education to vary by gender, unless the interaction effect is 0. I have estimated this model, with the following results: $b_0 = 26.29$, $b_1 = -.794$, $b_2 = -4.79$, $b_3 = .118(ns)$:

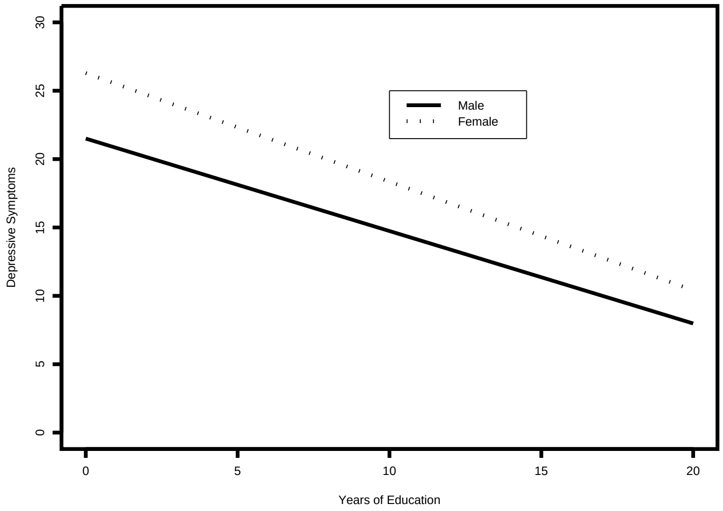

Figure 2. Depressive Symptoms by Years of Education and Gender (with Interaction)

Notice that in this plot, the lines for women and men appear to converge slightly as education increases. This implies that women gain more, in terms of reduction of depressive symptoms, from education than men do.

In this example, the interaction effect is not significant, and I didn't expect it to be. There is no theory that suggests an interaction between gender and education.

In the following example, I examine racial differences in depressive symptoms. The results for a model with race (white) only are: $b_0 = 17.14$, $b_{white} = -2.92$. In a second model, I examine the racial difference, net of education. The results for that model are: $b_0 = 25.62$, $b_{educ} = -.73$, $b_{white} = -1.85$:

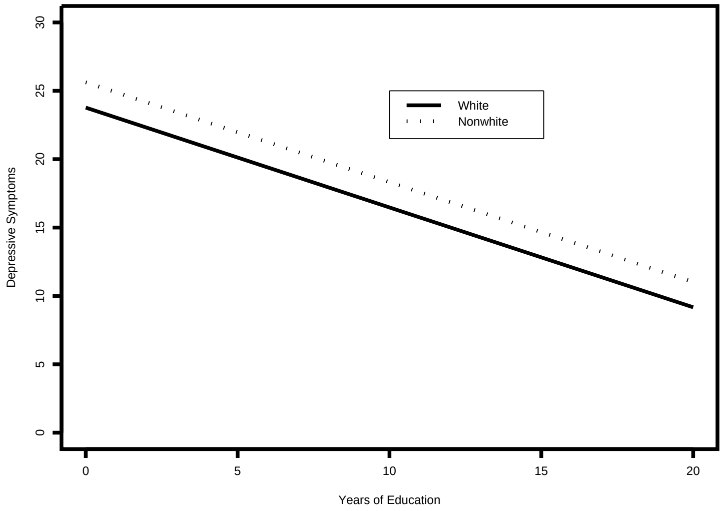

Figure 3. Depressive Symptoms by Years of Education and Race

The change in the coefficient for 'white' between the two models indicates that education accounts for part of the racial difference in symptoms. More specifically, it suggests that the large difference observed in the first model is in part attributable to compositional differences in educational attainment between white and nonwhite populations. The second model suggests that whites have lower levels of depressive symptoms than nonwhites, but that education reduces symptoms for both groups.

Theory might suggest that nonwhites (blacks specifically) get fewer returns from education, though, so in the next model, I include a *white* × *education* interaction term. Those results ($b_0 = 21.64$, $b_{educ} = -.39$, $b_{white} = 4.34$, $b_{ed\times w} = -.51$) yield a plot that looks like:

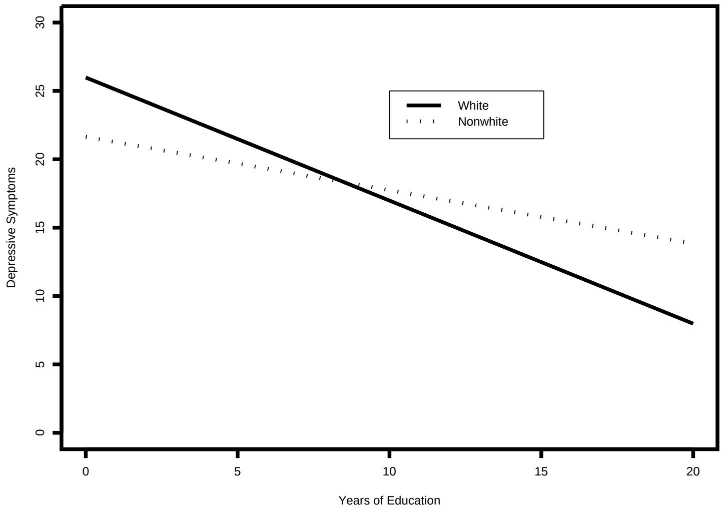

Figure 4. Depressive Symptoms by Years of Education and Gender (with Interaction)

Here, it is apparent that whites actually have greater depressive symptoms than nonwhites at very low levels of education, perhaps because nonwhites are more adaptive to adverse economic conditions than whites are. However, as education increases, symptoms decrease much faster for whites than for nonwhites. In fact, we could determine the precise level of education at which the lines cross, by setting the nonwhite and white results equal to each other and solving for Education:

$$\begin{aligned} White\ Symptoms &= Nonwhite\ Symptoms \\ 21.64 - .38E + 4.34 - .51E &= 21.64 - .39E \\ E &= 8.51 \end{aligned}$$

We may choose to get more elaborate. So far, I have alluded to the fact that these racial results may be attributable to differences between blacks and whites—we may not find the same results if we further subdivided our race variable, but this is an empirical question. Thus, I conducted an additional set of models in which I disaggregated the 'nonwhite' category into 'black' and 'other' categories, and I constructed interaction terms between education and the 'white' and 'other' dummy variables. I obtained the following results

Variable	Model 1	Model 2	Model 3
Intercept	17.4***	26.2***	22.8***
White	-3.18***	-2.32***	3.22#
Other	-.84	-1.53*	-2.22
Educ		-.74***	-.45***
E×W			-.45**
E×O			.09

What do these results suggest? It appears, based on Model 1, that whites have significantly fewer symptoms than blacks, but that 'others' don't differ much from blacks. The second model indicates that educational attainment is important, and, interestingly, the 'other' coefficient is now significant. Here we observe a suppressor relationship—educational differences between 'blacks' and 'others' masks the differences in symptoms between these races. This result implies that 'other' races have lower educational attainment than blacks. Model 3 confirms our earlier result—that education provides greater returns to whites. It also dispels the hypothesis that this interaction effect is unique to blacks (blacks and others appear to have similar patterns, given the nonsignificance of the coefficients for 'others.') It is important to note that the 'other' category is composed (in this study) primarily of Hispanics (who do, in fact, have lower educational attainment than blacks). If we were to further subdivide the race variable, we may find that heterogeneity within the 'other' category is masking important racial differences.

2.2 Interactions Between Continuous Variables

So far, we have been able to represent regression models with a single variable (continuous or dummies), and models with a single continuous variable and a set of dummies, and models with interactions between a single continuous variable and a dummy variable, with two-dimensional graphs. However, we will now begin considering models which can be represented graphically only in three dimensions. Just as a continuous variable may have an effect that varies across levels of a dummy variable, we may have continuous variables whose effects vary across levels of another continuous variable. These interactions become somewhat more difficult to perceive visually.

Below, I have shown a two variable model in which depressive symptoms were regressed on functional limitations (ADLs) and years of education. The regression function is a plane, rather than a line, as indicated in the figure.

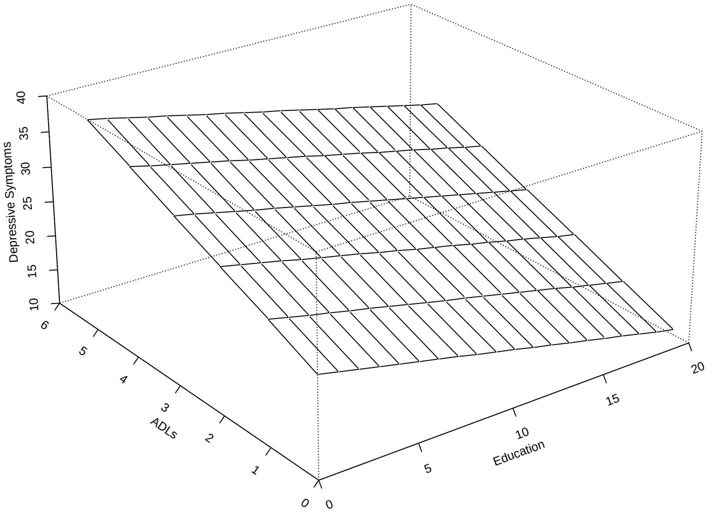

Figure 5. Depressive Symptoms by Years of Education and ADLs

Notice that the plane is tilted in both ADL and Education dimensions. This tells us that as we move down in education or up in ADLs, we get increases in depressive symptoms. Obviously, the lowest depressive symptoms occur for highly-educated unlimited individuals, and the greatest symptoms occur for low-educated highly-limited individuals. Suppose that we don't believe this pattern is linear in both dimensions, but rather that there may be some synergy between education and ADLs. For example, we might expect that the combination of physical limitation and low education to be far worse, in terms of producing depressive symptoms, than this additive model suggests. In that case, we may include an interaction term between education and ADL limitations.

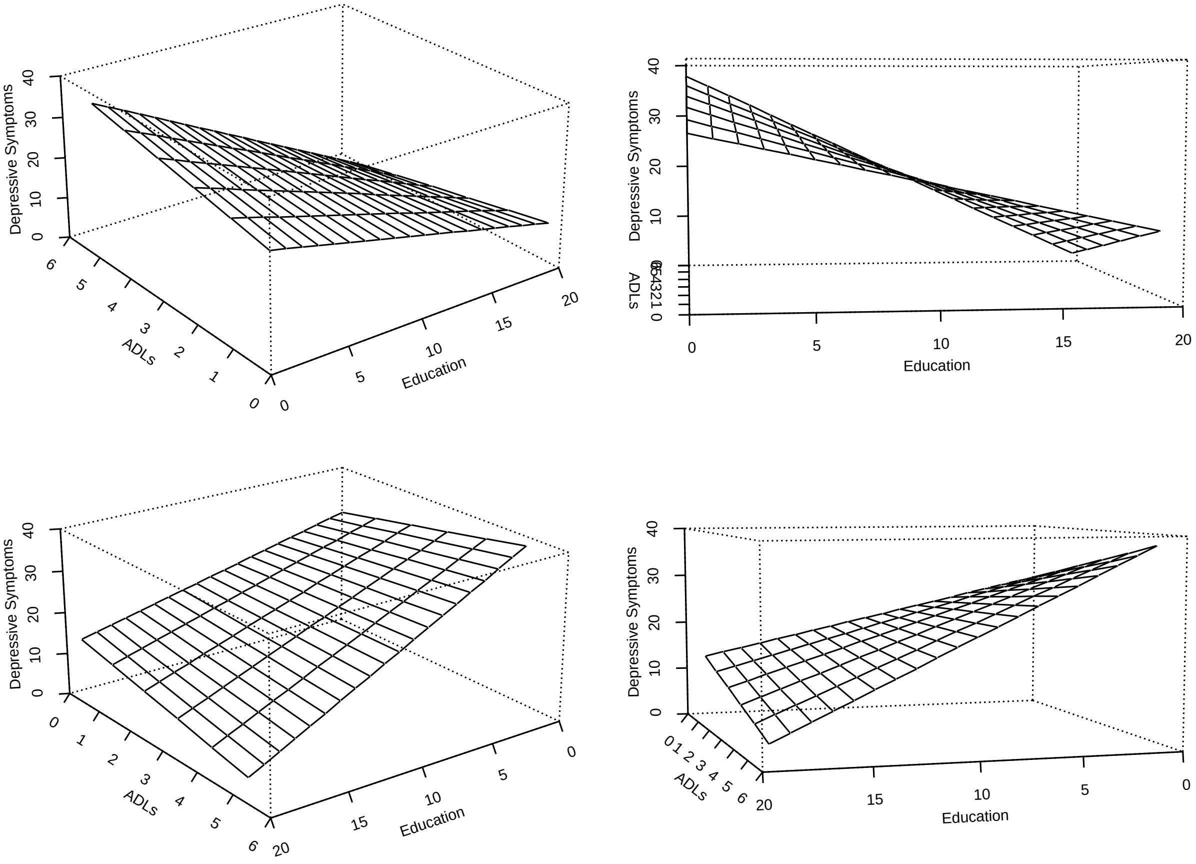

Figure 6. Depressive Symptoms by Years of Education and ADLs (with Interaction)

Now we are no longer dealing with a flat plane, but rather a twisted surface.

If you don't want to plot the model-predicted data, you can always algebraically factor the model to determine what the total effect of one variable is, but you may need to do this in both dimensions. For example, this model is:

$$\hat{Y} = b_0 + b_1 Education + b_2 ADLs + b_3(Education \times ADLs)$$

Thus, the total effect of education is:

$$Education\, Effect = b_1 + b_3 ADLs$$

Similarly, the total effect of ADLs is:

$$ADL\, Effect = b_2 + b_3 Education$$

If you want, you can plot these two-dimensional effects, but it is often useful to just plug in the numbers and observe what happens. I find the following table useful for making interpretations:

↓ XZ Interaction	Effect of X (+)	Effect of X (-)
(+)	Effect of X grows across levels of Z	Effect of X shrinks across levels of Z
(-)	Effect of X shrinks across levels of Z	Effect of X grows across levels of Z

In our example, the individual effect of education is negative, the effect of ADLs is positive, and the interaction effect is negative. This implies that the total effect of education increases across levels of disability, and that the total effect of disability decreases across levels of education. We can discern this from the graph above-it appears that at very high levels of education, adding ADL limitations increases depressive symptoms slower (in fact, it actually *reduces* symptoms) than it does at low levels of education. In the other dimension, it appears that adding years of education has a relatively large effect on reducing depressive symptoms among persons with 6 ADLs compared to the effect among persons with 0 ADLs. Substantively, these results suggest that physical limitation is particularly troubling for lower SES individuals, and that it is less troubling for persons with high SES. Certainly, we could attempt to pin down an explanation by including additional variables, and here we enter the realm of some complex interpretation. Whereas without interactions, if one variable (Z), when entered into a model, reduces the effect of another (X), we may argue that Z partially explains the effect of X. We can make similar interpretations with interactions. If, for example, we entered some measure of 'sense of control' into the above model, and the interaction effect disappeared or reversed, we have some evidence that differential feelings of control explains why functional limitations have less effect on depressive symptoms for higher-SES individuals. Sometimes, however, we may need an interaction to explain an interaction.

2.3 Limitations of Interactions

In theory, there are no limits to the number interaction terms that can be put into a model, and there are no limits on the degree of interaction. For example, you can construct three-way or even higher order interactions. But, there is a tradeoff, in terms of parsimony and interpretability. Multi-way interactions become very difficult to interpret, and they also require complex theories to justify their inclusion in a model. As an example from my own research on education and health, it is clear that there is a life-course pattern in the education-health relationship. That is, education appears to become more important across age. Hence, education interacts with age. It is also clear that education is becoming more important in differentiating health across birth cohorts-education has a stronger relationship to health for more recent cohorts than older birth cohorts, perhaps due to improvement in the quality of education (content-wise), or perhaps due to the greater role education plays today in getting a high-paying job. Taken together, these arguments imply a three-way interaction between age, birth cohort, and education. However, a simple three-way interaction would be very difficult to interpret.

A second difficulty with interactions occurs when all the variables that are thought to interact are dummy variables. For example, suppose from the previous example, age were indicated by a dummy variable representing being 50 years old or older (versus younger than 50), education were indicated by a dummy representing having a high school diploma (versus not), and cohort was indicated by a dummy representing having a birth date pre-WWII (versus post-WWII). A simple three-way interaction between these variables would ONLY take a value of 1 for persons who were a 1 on all three of these indicators. Yet, this may not be the contrast of interest.

There are no simple rules to guide you through the process of selecting reasonable interactions—it just takes practice and thoughtful consideration of the hypotheses you seek to address.

3 Nonlinear Transformations

We often find that the standard linear model is inappropriate for a given set of data, either because the dependent variable is not normally distributed, or because the relationship between the independent and dependent variables is nonlinear. In these cases, we may make suitable transformations of the independent and/or dependent variables to coax the dependent variable to normality or to produce a linear relationship between X and Y. Here, we will first discuss the case in which the dependent variable is not normally distributed.

3.1 Transformations of Dependent Variables

A dependent variable may not be normally distributed, either because it simply doesn't have a normal-bell shape, or because its values are bounded, creating a skew. In other cases, it is possible that the variable has a nice symmetric shape, but that the interval on which the values of the variable exist is narrow enough to produce unreasonable predicted values (e.g., the dependent variable is a proportion). In terms of inference for parameters, while a nonnormal dependent variable doesn't guarantee a nonnormal distribution of errors, it generally will. As stated previously, this invalidates the likelihood function, and it also invalidates the derivation of the sampling distributions that provide us with the standard errors for testing. Thus, we may consider a transformation to normalize the dependent variable.

There are 4 common transformations that are used for dependent variables, including the logarithmic, exponential, power, and logistic transformations:

Transformation	New D.V.	New Model
Logarithmic	$Z = ln(Y)$	$Z = Xb$
Exponential	$Z = exp(Y)$	$Z = Xb$
Power	$Z = Y^p, p \neq 0, 1$	$Z = Xb$
Logistic	$Z = \frac{ln(Y)}{(1-ln(Y))}$	$Z = Xb$

Notice in all cases, the new model is the transformed dependent variable regressed on the regressors. This highlights that there will be a change in the interpretation of the effects of X on Y. Specifically, you can either interpret the effects of X on Z, or you can invert the transformation after running the model and computing predicted scores in Z units.

Each of these transformations has a unique effect on the distribution of Y. The logarithmic transformation, as well as power transformations in which $0 < p < 1$, each reduce right skew. Why? Because the $log_b X$ function is a function whose return value is the power to which b must be raised to equal X. For example, $log_{10} 100 = 2$. Below shows the pattern for log base 10:

X	Log X
1	0
10	1
100	2
1000	3
10000	4

You can see that, although X is increasing by powers of 10, $logX$ is increasing linearly. We tend to use the natural log function, rather than the base 10 function, because it has nice properties. The base of the natural logs is $e \approx 2.718$. This transformation will tend to pull large values of Y back toward the mean, reducing right skewness. By the same token, power transformations ($0 < p < 1$) will have a similar effect, only such transformations generally aren't as strong as the log transformation.

Recognize that changing the distribution of Y will also have an effect on the relationship between X and Y. In the listing above of the logarithms of the powers of 10, the transformation changes the structure of the variable, so that an X variable which was linearly related to powers of 10 will now be nonlinearly related to the log of these numbers.

The log and power transformations are used very often, because we encounter right skewness problems frequently. Any variable that is bounded on the left by 0 will tend to have right skewed distributions, especially if the mean of the variable is close to 0 (e.g., income, depressive symptoms, small attitude scales, count variables, etc.).

When distributions have left skewness, we may use other types of power transformations in which $p > 1$, or we may use the exponential transformation. By the same token that roots and logs of variables pull in a right tail of a distribution, the inverse functions of roots and logs will expand the right tail of a distribution, correcting for a left skew.

What if $p < 0$ in a power transformation? This does two things. First, it inverts the distribution-what were large values of Y would now be small values after the transformation (and vice versa). Then, it accomplishes the same tail-pulling/expanding that power transformations in which p is positive do. We don't often see these types of transformations, unless we are dealing with a variable in which interest really lay in the inverse of it. For example, we may take rates and invert them to get time-until-an event. The main reason that we don't see this type of transformation frequently is that it doesn't affect skewness any more than a positive power transformation would.

We need to be careful when applying the log and power transformations. We can't take even roots when a variable takes negative values. Similarly, we can't take the logarithm of 0. When we are faced with these problems, we may simply add a constant to the variable before transforming it. For example, with depressive symptoms, we may wish to add 1 to the variable before taking the log to ensure that our logarithms will all exist.

The final transformation we will discuss is the logistic transformation. The logistic transformation is useful when the outcome variable of interest is a proportion-a number bounded on the $[0, 1]$ interval. The distribution of such a variable might very well be symmetric, but the bounding poses a problem, because a simple linear model will very likely predict values less than 0 and greater than 1. In such a case, we may take:

$Z = log\frac{Y}{1-Y}$. This transformation yields a boundless variable, because $Z = -\infty$ when Y approaches 0, and $Z = +\infty$ when Y approaches 1. Of course, because of a division by 0 problem, and because the log of 0 is undefined, something must be done to values of 0 or 1.

3.2 Transformations of Independent Variables

Whereas we may be interested in transforming a dependent variable so that the distribution of the variable becomes normal, there is no such requirement for independent variables-they need not be normally distributed. However, we often perform transformations on independent variables in order to linearize the relationship between X and Y. Before we discuss methods for linearizing a relationship between X and Y, we should note that what makes the 'linear model' linear is that the effects of the parameters are additive. Stated another way, the model is linear in the parameters. We may make all the transformations of the variables themselves that we like, and the linear model is still appropriate so long as the model parameters can be expressed in an additive fashion. For example:

$$Y = exp(b_0 + b_1X)$$

is a linear model, because it can be re-expressed as:

$$ln(Y) = b_0 + b_1X.$$

However, the model: $Y = b_0^{b_1X}$ is not a linear model, because it cannot be re-expressed in an additive fashion.

The same transformations that we used for altering the distribution of the dependent variable can also be used to capture nonlinearity in a regression function. The most common such transformations are power transformations, specifically squaring and cubing the independent variable. When we conduct such transformations on independent variables, however, we generally include both the original variable in the model as well as the transformed variables—just as when we included interaction terms, we included the original 'main effects.' (note: this is a matter of convention, however, and not a mathematical requirement!) Thus, we often see models like:

$$\hat{Y} = b_0 + b_1X + b_2X^2 + b_3X^3 \dots$$

Such models are called 'polynomial regression models,' because they include polynomial expressions of the independent variables. These models are good for fitting regression models when there is relatively mild curvature in the relationship between X and Y. For example, it is well-known that depressive symptoms tend to decline across young adulthood, bottoming out after retirement, before rising again in late life before death. This suggests a quadratic relationship between age and depressive symptoms rather than a linear relationship. Thus, I conducted a regression model of depressive symptoms on age, and then a second model with an age-squared term included:

Variable	Model 1	Model 2
Intercept	18.47***	23.34***
Age	-.085***	-.322***
Age^2		.0025***

The first model indicates that age has a negative effect—every year of age contributes to a .085 symptom reduction. The second model reveals that age reduces symptoms, but also that each year of age contributes a smaller and smaller reduction. Ultimately, when age is large enough, age will actually begin to lead to increases in symptoms. We can see this by taking the first derivative of the regression function with respect to age, setting it equal to 0, and solving to find the age at which age begins to raise symptoms:

$$\hat{Y} = b_0 + b_1 Age + b_2 Age^2.$$

The derivative is:

$$\frac{\partial \hat{Y}}{\partial Age} = b_1 + 2b_2 Age.$$

Solving for the age at which the effect reaches a minimum is:

$$Age_{min} = \frac{-b_1}{2b_2}$$

In this case, if we substitute our regression estimates in, we find the age is 64.4. The plot below confirms this result.

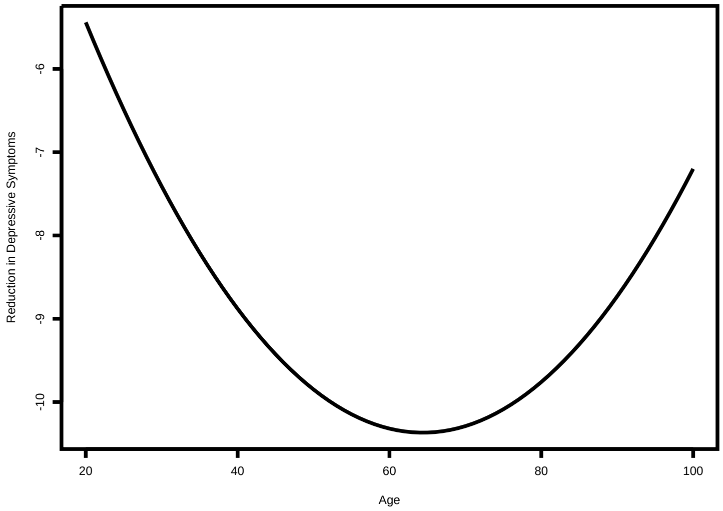

Figure 7. Net Effect of Age on Depressive Symptoms: Second Degree (Quadratic) Polynomial Regression

In the plot, it is clear that the effect of age reaches a minimum in the mid 60's and then becomes positive afterward.

We can use polynomial regression functions to capture more than a single point of inflection. The rule is that we can capture $k-1$ points of inflection with a polynomial regression of degree k. As an example, I reconducted the above regression model with an age-cubed term and got the following results:

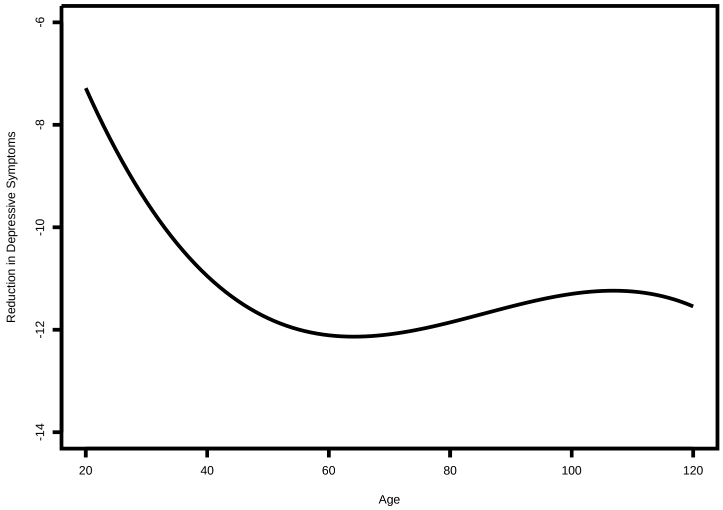

Figure 8. Net Effect of Age on Depressive Symptoms: Third Degree Polynomial Regression

In this case, it looks like the effect of age bottoms out just before retirement age, and that the effect of age is positive but relatively trivial beyond retirement. Additionally, it appears that if one lives past 100 years of age, age again contributes to reducing symptoms (a second point of inflection). Of course the age-cubed term was nonsignificant, though, and really shouldn't be interpreted.

I should note that as you add polynomial terms, you can ultimately almost perfectly fit the data (and you can perfectly fit it, if there is only one person at each value of X). However, the tradeoff of a perfect fit is a loss of parsimony.

3.3 Dummy Variables and Nonlinearity

Sometimes, nonlinearity is severe enough or odd-patterned enough that we can't easily capture the nonlinearity with a simple transformation of either (or both) independent or dependent variables. In these cases, we may consider a spline function. Splines can be quite complicated, but for our purposes, we can accomplish a spline using dummy variables and polynomial expressions. In this case, we sometimes call the model a 'piecewise regression.' Imagine that the regression function is linear for values of X less than C and also linear for values above C:

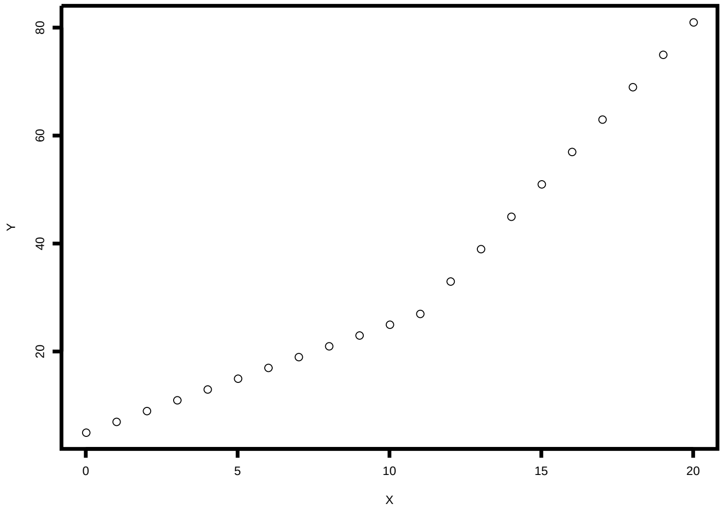

Figure 9. Data That are Linear in Two Segments

In this figure, the relationship between X and Y is clearly linear, but in sections. Specifically, the function is:

$$Y = \begin{cases} 5 + 2X & iff\ X < 11 \\ 27 + 6(X - 11) & iff\ X > 10 \end{cases}$$

If we ran separate regression models for $X < 11$ and $X > 10$ (e.g., constructed two samples), we would find a perfectly linear fit for both pieces. However, each section would have a different slope and intercept. One solution is to estimate a second-degree polynomial regression model ($\hat{Y} = b_0 + b_1X + b_2X^2$):

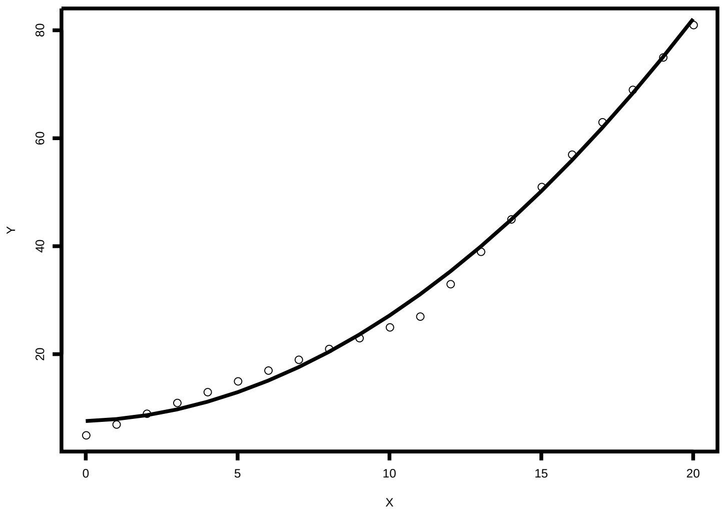

Figure 10. Quadratic Fit to Two-Segment Data

This model offers a fairly decent fit in this case, but it tends to underpredict and overpredict systematically, which produces serial correlation of the errors (violating the independence assumption). Plus, this data could be predicted perfectly with another model.

We could construct a dummy variable indicating whether X is greater than 10, and include this along with X in the model:

$$\hat{Y} = b_0 + b_1X + b_2(I(X > 10)).$$

This model says $\hat{Y}$ is $b_0 + b_1X$, when $X < 11$, and $(b_0 + b_2) + b_1X$ when $X > 10$. This model doesn't work, because only the intercept has changed—the slope is still the same across pieces of X. But, we could include an interaction term between the dummy variable and the X variable itself:

$$\hat{Y} = b_0 + b_1X + b_2(I(X > 10)) + b_3(I \times X).$$

This model says $\hat{Y} = b_0 + b_1X$, when $X < 11$, and $(b_0 + b_2) + (b_1 + b_3)X$, when $X > 10$. We have now perfectly fit the data, because we have allowed both the intercepts and slopes to vary across the 'pieces' of the data:

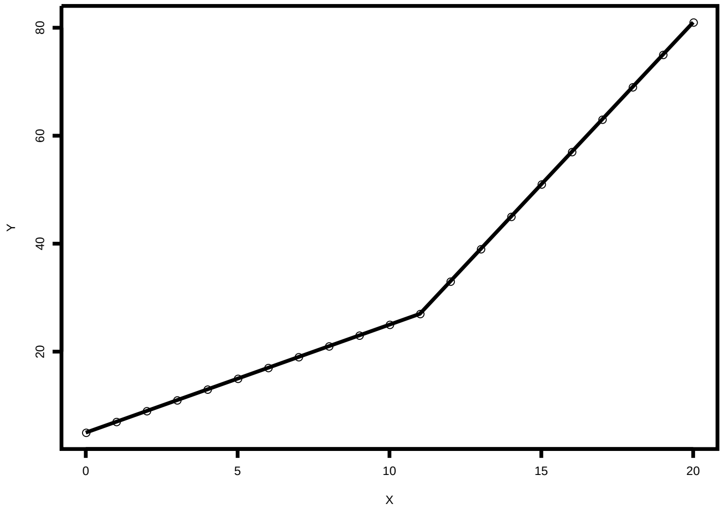

Figure 11. Spline (with Single Fixed Knot and Degree=1) or Piecewise Regression Model: Observed and Predicted Scores

This model can be called either a piecewise regression, or a spline regression. Notice the title of the figure includes “degree 1” and “1 knot.” The ‘knots’ are the number of junctions in the function. In this case, there are 2 regression functions joined in one place. The ‘degree’ of the spline is 1, because our function within each regression is a polynomial of degree 1. We can, of course include quadratic functions in each piece, and as we include more and more of them, the function becomes more and more complex because of the multiple interaction terms that must be included. Technically, a spline function incorporates the location of the knot as a parameter to be estimated (and possibly the number of knots, as well).

3.4 Final Note

For those who are interested in constructing a cognitive map of statistical methods, here are a few tidbits. First, conducting the log transformation on a dependent variable leaves us with one form of a poisson regression model, which you would discuss in more depth in a class on generalized linear models (GLM). Also, the logistic transformation is used in a model called ‘logistic regression,’ which you would also discuss in a GLM class. However, the logistic transformation we are discussing, while the same as the one used in logistic regression, is not applied to the same type of data. In logistic regression, the outcome variable is dichotomous (or polytomous). Finally, the piecewise/spline regression discussed above is approximately a special case of a neural network model.