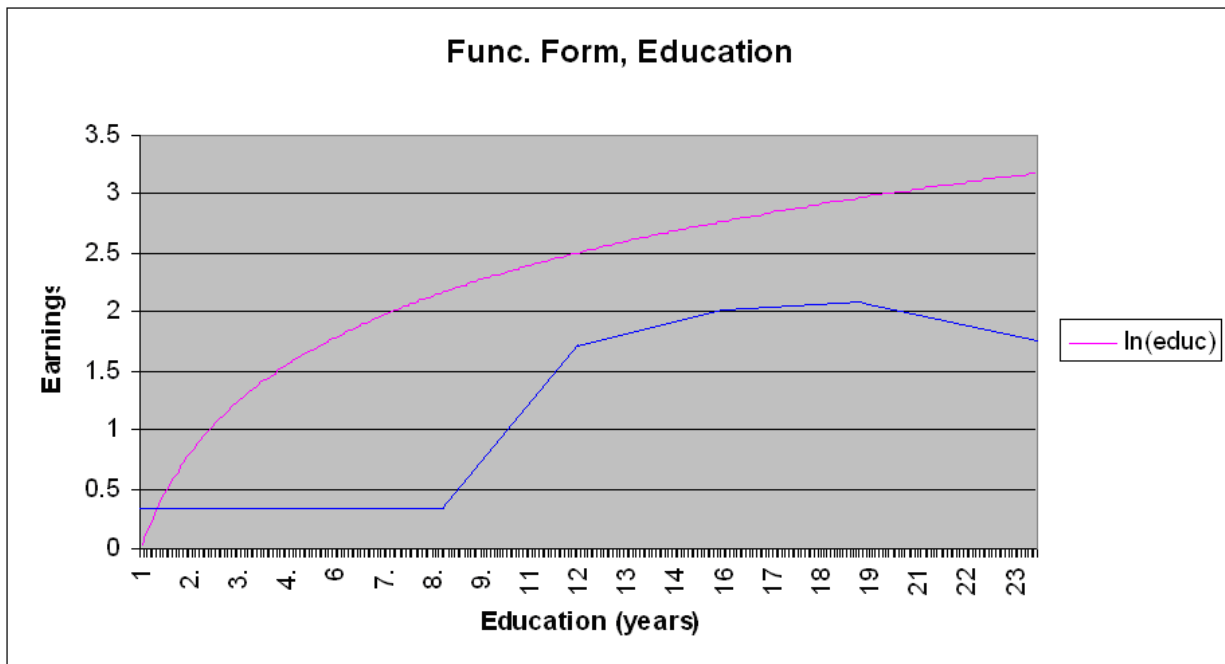


**PAD 705 Handout: Piecewise Linear Regression**

Earlier in the term we discussed at length the use of non-linear functional forms as a way to better fit the data. All the forms we discussed – linear, quadratic, natural log, etc. – were “smooth and continuous” (see the graph below).



In other words, the line has no kinks, no gaps, and no endpoints. In the graph above, the natural log function is both smooth and continuous. However, there are many phenomena where this assumption may be erroneous for some of the independent variables. In many cases, there are “milestones” that cause relationships to change. The most celebrated one is in education. Education in the United States is broken into several institutionally and functionally distinct periods of time: Elementary education (grades 1 to grade 8); high school (to grade 12); undergraduate (to grade 16); graduate (to grade 19); and doctoral (to grade 23 or more). If we looked at the relationship between earnings and education, it may be that earnings do *not* always climb at a decreasing rate, as the natural log shape implies. Instead, there may be periods of virtually no relationship (elementary), rapid increase (high school); slower but still positive increase (undergraduate and graduate study); and decrease (doctoral-level work, unfortunately). If we believe that a smooth, continuous functional form is not supported by theory, past research, or intuition, we need to try another shape. The goal is to see if we can increase the  $R^2$  by better specifying the form.

### *Using a Piecewise Linear Specification*

If you look carefully at the graph on the first page, note that the lines that are connected, with the “join” points for each segment corresponding to the end of a period of education. Notice also that each has both a different slope and different intercept. Our specification process must result in joined line segments, with the points where they connect being at the end of one range and the beginning of another.

Notice another thing: a person who drops out of high school after 10<sup>th</sup> grade has completed only two years of high school, but all 8 years of elementary school. Somehow, we need to have this person’s eight years of elementary education “count” toward the slope and intercept for the “elementary” segment and the person’s two years of high school education count toward the “high school” segment.

To accomplish all of these things, we will use a set of dummy variables and “transformations” of the education variable. First, we need to specify which members of the data set have at least some high school education, at least some undergraduate education, at least some graduate education, and at least some doctoral education. We will use dummy variables for this purpose. Second, we will take into account years in each range above elementary education by subtracting the cutoff value for the *previous* segment from the person’s educational attainment. For instance, the last year of elementary education is 8<sup>th</sup> grade. To figure out how many years of education the person attained (if any) above elementary school, we subtract 8 from their educational attainment. To figure out how many years of education the person attained above high school, we will subtract 12 from their educational attainment.

Finally, to keep a person from having a negative number in the “transformed” variable because they have not attended high school, undergraduate education, graduate education, or doctoral study, we will multiply each transformation of education by a dummy for the segment. For instance, if a person dropped out of high school, they have completed no more than 11 years of education. The value of undergraduate transformation is given by the calculation (education – 12), which will be negative for this person. To make sure it is set to zero for people who have no undergraduate study, we will multiply the transformed education variable by the dummy variable for “some undergraduate” education, which will be zero for a person who did not complete high school.

After all of these calculations, we will end up with four dummy variables, four transformations, and four “dummy times transformation” variables. The regression will include a constant, the original education variable, and all four “dummy times transformation” variables.

### *Example: Problem Set #2*

In Problem Set #2 we analyzed the relationship between log hourly wage and several independent variables, including education. We will use the `cps83.dta` data set to test a piecewise linear regression specification on education. First, let’s look at a specification where education enters linearly but experience is modeled with a quadratic specification.

```
. reg lhwage female pcfemale exper exper2 NE NC SO union yrseduc
```

Source	SS	df	MS	Number of obs = 801		
Model	71.6026235	9	7.95584705	F( 9, 791)	=	48.75
Residual	129.093544	791	.163202963	Prob > F	=	0.0000
-----				R-squared	=	0.3568
-----				Adj R-squared	=	0.3495
Total	200.696167	800	.250870209	Root MSE	=	.40398

lh wage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.2095853	.0382031	-5.49	0.000	-.2845766	-.1345939
pcf female	-.0018932	.0005926	-3.19	0.001	-.0030565	-.0007299
exper	.0373371	.0038002	9.82	0.000	.0298774	.0447968
exper2	-.0005724	.0000802	-7.14	0.000	-.0007298	-.000415
NE	-.0526285	.0432121	-1.22	0.224	-.1374525	.0321955
NC	-.0712214	.0405213	-1.76	0.079	-.1507633	.0083205
SO	-.1514514	.0388467	-3.90	0.000	-.2277061	-.0751966
union	.1045543	.0325778	3.21	0.001	.0406052	.1685035
yrseduc	.0747347	.0055379	13.50	0.000	.063864	.0856054
_cons	.8812849	.0895787	9.84	0.000	.7054447	1.057125

Age was not included because it is highly correlated with experience (i.e., multicollinearity is a problem). The adjusted  $R^2$  is .3495 – good but not great.

Now let's start creating the variables we need for this specification. First, we will create dummies for the periods of education. The data has an education top-code of 18 years, so we will have four segments: elementary, high school, undergraduate, and graduate. We will assume that everyone has at least some elementary education, so we need dummies for the other three:

```
. gen SomeHS = ( yrseduc>8)
. gen SomeUG = ( yrseduc>12)
. gen SomeGrad = ( yrseduc>16)
```

Next, we will create the educational transformations. Again, we will use three new variables:

```
. gen EOverElem = ( yrseduc-8)
. gen EOverHS = ( yrseduc-12)
. gen EOverUG = ( yrseduc-16)
```

Now we need to interact the dummies and the transformations to make sure those who have less than the floor levels of education in high school, undergraduate study, and graduate study are set to zero:

```
. gen transHS = SomeHS* EOverElem
. gen transUG = SomeUG* EOverHS
. gen transGrad = SomeGrad* EOverUG
```

We are now ready to run the regression:

```
. reg lhwage female pcfemale exper exper2 NE NC SO union yrseduc transHS transUG
transGrad
```

Source	SS	df	MS	Number of obs = 801		
Model	74.151293	12	6.17927442	F( 12, 788) =	38.48	
Residual	126.544874	788	.160589942	Prob > F =	0.0000	
Total	200.696167	800	.250870209	R-squared =	0.3695	
				Adj R-squared =	0.3599	
				Root MSE =	.40074	

lhwage	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	-.1987573	.0380425	-5.22	0.000	-.273434	-.1240806
pcfemale	-.0020473	.0005898	-3.47	0.001	-.0032051	-.0008895
exper	.0382	.0037925	10.07	0.000	.0307554	.0456447
exper2	-.0005925	.0000806	-7.35	0.000	-.0007507	-.0004343
NE	-.0408261	.0430852	-0.95	0.344	-.1254014	.0437492
NC	-.0554558	.0404021	-1.37	0.170	-.1347643	.0238527
SO	-.1364458	.0387429	-3.52	0.000	-.2124973	-.0603944
union	.1085255	.032362	3.35	0.001	.0449996	.1720514
yrseduc	-.0032198	.0244959	-0.13	0.895	-.0513047	.0448651
transHS	.0888109	.0346236	2.57	0.011	.0208456	.1567762
transUG	.0121536	.0209923	0.58	0.563	-.0290537	.053361
transGrad	-.0743834	.0369339	-2.01	0.044	-.1468838	-.0018829
_cons	1.422323	.1827126	7.78	0.000	1.063662	1.780984

Two of the four education-related variables are statistically significant – the one for the high school segment and the one for the graduate segment. In essence, the segments for elementary education and undergraduate education have a zero slope. Does this specification do better than our original one, where education entered linearly? We can answer this question using an F test, because the only difference between our first and second regressions is the inclusion of the “transformation times dummy” variables?

```
. test transGrad transUG transHS

( 1) transGrad = 0.0
( 2) transUG = 0.0
( 3) transHS = 0.0

F( 3, 788) = 5.29
Prob > F = 0.0013
```

So this specification is better: the  $R^2$  is .3599 (versus .3495) and the new variables are jointly significant. (Unhappily for us, graduate education again seems to be related to decreased earnings.)

The final issue to examine carefully is the slope and intercept for each segment. Let’s begin by looking at the regression equation (to simplify, I will drop the “i” subscript):

$$\begin{aligned}
 \text{lhwage} = & \beta_0 + \beta_1 \text{yrseduc} + && \text{] Elementary education only} \\
 & \beta_2 (\text{yrseduc} - 8) * D_{\text{some\_high\_school}} + && \text{] Some High school, but no more} \\
 & \beta_3 (\text{yrseduc} - 12) * D_{\text{some\_undergraduate}} + && \text{] Some Undergraduate, but no more} \\
 & \beta_4 (\text{yrseduc} - 16) * D_{\text{some\_graduate}} + \varepsilon && \text{] Some Graduate}
 \end{aligned}$$

The key to interpreting these findings is to realize what segments apply to a particular observation and then to multiply the estimated coefficients by the components of the transformation. For instance, the coefficient  $\beta_2$  must be multiplied by  $\text{yrseduc}$  and 8. The slope on the line for those with some high

