Example 4: General Linear Models with Multiple Fixed Effects of a Single Conceptual Predictor in SAS and STATA

The data for this example were selected from the 2012 General Social Survey dataset featured in Mitchell (2015); these data were also used for examples 1, 2, and 3. This example will use general linear models to predict a single quantitative outcome (annual income) when multiple fixed effects are needed to describe a predictor's relationship to the outcome: for categorical predictors with more than two categories (3-category working class), for quantitative predictors with nonlinear effects (quadratic years of age or piecewise years of education), or for testing the assumption of an single linear slope for ordinal predictors (5-category happiness).

SAS Syntax for Importing and Preparing Data for Analysis:

```
* Paste in the folder address where "GSS_Example.xlsx" is saved after = before ;
%LET filesave= \\Client\C:\Dropbox\21SP_PSQF6242\PSQF6242_Example4;

* IMPORT GSS_Example.xlsx data using filesave reference and exact file name;
* from the Excel workbook in DATAFILE= location from SHEET= ;
* New SAS file is in "work" library place with name "Example4";
* "GETNAMES" reads in the first row as variable names;
* DBMS=XLSX (can also use EXCEL or XLS for .xls files);
PROC IMPORT DATAFILE="&filesave.\GSS_Example.xlsx"
            OUT=work.Example4 DBMS=XLSX REPLACE;
     SHEET="GSS_Example";
     GETNAMES=YES;
RUN;
* Create formats: set of value labels for categorical variables;
PROC FORMAT;
     VALUE Fclass 1="1.Lower" 2="2.Middle" 3="3.Upper";
     VALUE Fhappy 1="1.Unhappy" 2="2.Neither" 3="3.Fairly Happy"
                  4="4.Very Happy" 5="5.Completely Happy";
RUN;
* All data transformations must happen inside a DATA+SET combo to know where to use them;
* Here is how to make a new variable: new = old;
DATA work.Example4; SET work.Example4;
* Label variables and apply value formats for variables used below;
* LABEL name=      "name: Descriptive Variable Label";
  LABEL workclass= "workclass: 3-Category Working Class"
        age=       "age: Years of Age"
        educ=      "educ: Years of Education"
        happy=     "happy: 5-Category Happy Rating"
        income=    "income: Annual Income in 1000s";
* Apply value labels created above: name Format.;
  FORMAT workclass Fclass. happy Fhappy.;
* Select cases complete on all variables to be used;
  WHERE NMISS(income,workclass,age,educ,happy)=0;
RUN;
* Now dataset work.Example4 is ready to use;
```

STATA Syntax for Importing and Preparing Data for Analysis:

```
// Paste in the folder address where "GSS_Example.xlsx" is saved between " "
   global filesave "\\Client\C:\Dropbox\21SP_PSQF6242\PSQF6242_Example4"

// IMPORT GSS_Example.xlsx data using filesave reference and exact file name
// To change all variable names to lowercase, remove "case(preserve")
   clear // Clear before means close any open data
   import excel "$filesave\GSS_Example.xlsx", case(preserve) firstrow clear
// Clear after means re-import if it already exists (if need to start over)

// Create formats: set of value labels for categorical variables
   label define Fclass    1 "1.Lower" 2 "2.Middle" 3 "3.Upper"
   label define Fhappy    1 "1.Unhappy" 2 "2.Neither" 3 "3.Fairly Happy" ///
                            4 "4.Very Happy" 5 "5.Completely Happy"
```

```
// Label variables and apply value formats for variables used below
// label variable name      "name: Descriptive Variable Label"
   label variable workclass "workclass: 3-Category Working Class"
   label variable age       "age: Years of Age"
   label variable educ      "educ: Years of Education"
   label variable happy     "happy: 5-Category Happy Rating"
   label variable income    "income: Annual Income in 1000s"
// Apply value labels created above: name Format
   label values workclass Fclass
   label values happy Fhappy
// Select cases complete on variables of interest
   egen nmiss = rowmiss(income workclass age educ happy)
   drop if nmiss>0
// Now dataset is ready to use
```

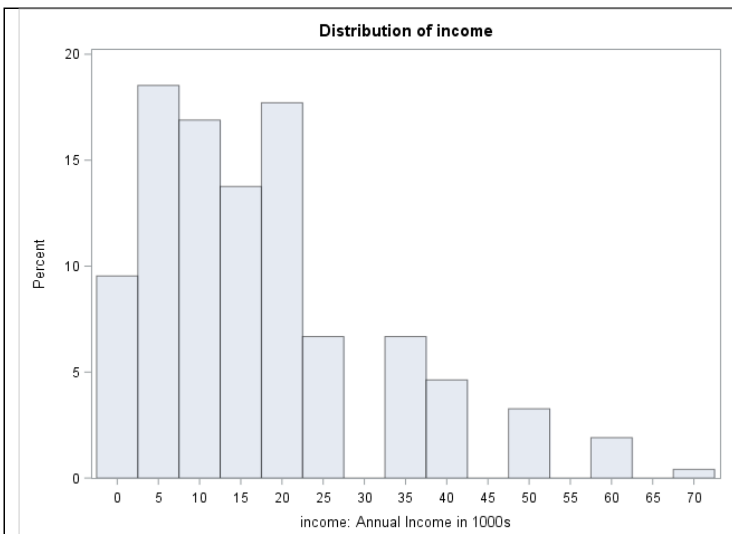
Syntax and SAS Output for Data Description and Empty Model for Income:

```
TITLE "SAS Descriptive Statistics for Quantitative income";
PROC MEANS NDEC=3 NOLABELS N MEAN STDDEV VAR MIN MAX DATA=work.Example4;
     VAR income;
RUN; TITLE;

display as result "STATA Descriptive Statistics for Quantitative income"
format income %5.3f
summarize income, format detail // detail to get variance
```

Analysis Variable: income

N	Mean	Std Dev	Variance	Minimum	Maximum
734	17.303	13.792	190.209	0.245	68.600

```
* Histograms to visualize quantitative variables;
* NOPRINT spares the rest of the results I do not want right now;
TITLE "SAS Histogram of Quantitative income";
PROC UNIVARIATE NOPRINT DATA=work.Example4;
     VAR income;
     HISTOGRAM income / MIDPOINTS=0 TO 70 BY 5;
RUN; QUIT; TITLE;

display "STATA Histogram of Quantitative income"
histogram income, percent discrete width(5) start(0)
```



Note that annual income in 1000s appears positively skewed (perhaps in part due to the logical lower-bound of 0).

But the real question is, what will its residuals look like after the model predictors are included? It's those residuals that are supposed be normal, not the original outcome variable. In stats language, this means that the distribution of y_i does not have to be normal *marginally*, but it should be normal *conditionally* (i.e., the e_i residuals after accounting for the model predictors should have a normal distribution) in order for the standard errors (and thus *p*-values) to be believable. Otherwise, you may need a different model than the GLM!

Empty Model to Predict Income: $Income_i = \beta_0 + e_i$

```
TITLE "SAS GLM Empty Model Predicting Income";
PROC GLM DATA=work.Example4 NAMELEN=100;
     MODEL income = / SOLUTION ALPHA=.05 CLPARM;
RUN; QUIT; TITLE;

display "STATA GLM Empty Model Predicting Income"
regress income, level(95)
```

NAMELEN extends printing of variable names; MODEL y = x / options (no x predictors so far); CLPARM provides confidence intervals (at chosen alpha level)

I am using STATA's "regression" procedure because it appears to be a general GLM version.

```
SAS GLM Empty Model Predicting Personal Income

                                          Sum of
Source                      DF           Squares     Mean Square    F Value    Pr > F
Model                        1       219751.8721     219751.8721    1155.32    <.0001
Error                      733       139423.2319        190.2090
Uncorrected Total          734       359175.1040

R-Square     Coeff Var      Root MSE    income Mean
0.000000      79.70716      13.79163       17.30287

                                    Standard
Parameter          Estimate            Error    t Value    Pr > |t|      95% Confidence Limits
Intercept       17.30287466       0.50905834      33.99      <.0001     16.30348846   18.30226086 Beta0
```

Mean Square Error (Mean Square **Residual** in STATA) gives the residual variance = 190.21. In the empty model this is ALL the outcome variance. No variance has been explained by predictors yet (R^2 = 0).

Syntax and SAS Output for 3-Category Working Class Predicting Income:

```
TITLE "SAS Descriptive Statistics for Categorical workclass";
PROC FREQ DATA=work.Example4;
     TABLE workclass;
RUN; TITLE;

display "STATA Descriptive Statistics for Categorical workclass"
tabulate workclass
```

```
             workclass: 3-Category Working Class
                                        Cumulative    Cumulative
workclass    Frequency     Percent      Frequency       Percent
----------------------------------------------------------------
 1.Lower          436       59.40            436         59.40
 2.Middle         278       37.87            714         97.28
 3.Upper           20        2.72            734        100.00
```

```
* SAS code to create dummy-coded binary predictors;
DATA work.Example4; SET work.Example4;
     LvsM=.; LvsU=.; * Make two new empty variables;
     IF workclass=1 THEN DO; LvsM=0; LvsU=0; END; * Replace each for lower;
     IF workclass=2 THEN DO; LvsM=1; LvsU=0; END; * Replace each for middle;
     IF workclass=3 THEN DO; LvsM=0; LvsU=1; END; * Replace each for upper;
     LABEL LvsM="LvsM: Low=0 vs Mid=1 Class"
           LvsU="LvsU: Low=0 vs Upp=1 Class";
RUN;

// STATA code to create dummy-coded binary predictors
   gen LvsM=. // Make two new empty variables
   gen LvsU=.
   replace LvsM=0 if workclass==1 // Replace each for lower
   replace LvsU=0 if workclass==1
   replace LvsM=1 if workclass==2 // Replace each for middle
   replace LvsU=0 if workclass==2
   replace LvsM=0 if workclass==3 // Replace each for upper
   replace LvsU=1 if workclass==3
```

Group (N = 734)	LvsM	LvsU
1. Low (n = 436)	0	0
2. Mid (n = 278)	1	0
3. Upp (n = 20)	0	1

```
label variable LvsM "LvsM: Low=0 vs Mid=1 Class"
label variable LvsU "LvsU: Low=0 vs Upp=1 Class"
```

Model with workclass via two dummy-coded predictors:

$$Income_i = \beta_0 + \beta_1(LvsM_i) + \beta_2(LvsU_i) + e_i$$

Predicted $\hat{y}_i = \beta_0 + \beta_1(LvsM_i) + \beta_2(LvsU_i)$

Low Mean: $\hat{y}_L = \beta_0 + \beta_1(0) + \beta_2(0) = \beta_0$ ← fixed effect #1
Mid Mean: $\hat{y}_M = \beta_0 + \beta_1(1) + \beta_2(0) = \beta_0 + \beta_1$ ← linear combination
Upp Mean: $\hat{y}_U = \beta_0 + \beta_1(0) + \beta_2(1) = \beta_0 + \beta_2$ ← linear combination

Diff of Low vs Mid: $(\beta_0 + \beta_1) - (\beta_0) = \beta_1$ ← fixed effect #2
Diff of Low vs. Upp: $(\beta_0 + \beta_2) - (\beta_0) = \beta_2$ ← fixed effect #3
Diff of Mid vs Upp: $(\beta_0 + \beta_2) - (\beta_0 + \beta_1) = \beta_2 - \beta_1$ ← linear combination

```
TITLE "SAS GLM Predicting Income from 3-Category workclass";
PROC GLM DATA=work.Example4 NAMELEN=100; * PLOTS(UNPACK)=DIAGNOSTICS;
     MODEL income = LvsM LvsU / SOLUTION ALPHA=.05 CLPARM;
     * Ask for predicted income per group and group differences;
     ESTIMATE "Low Mean"    intercept 1 LvsM  0 LvsU 0;
     ESTIMATE "Mid Mean"    intercept 1 LvsM  1 LvsU 0;
     ESTIMATE "Upp Mean"    intercept 1 LvsM  0 LvsU 1;
     ESTIMATE "Low vs Mid Diff"         LvsM  1 LvsU 0;
     ESTIMATE "Low vs Upp Diff"         LvsM  0 LvsU 1;
     ESTIMATE "Mid vs Upp Diff"         LvsM -1 LvsU 1;
     * Save requested estimates as SAS dataset to do math on them;
     ODS OUTPUT Estimates=work.ClassEstimates;
RUN; QUIT; TITLE;

display "STATA GLM Predicting Income from 3-Category workclass"
regress income c.LvsM c.LvsU, level(95)
// Ask for predicted income per group and group differences
lincom _cons*1 + c.LvsM*0  + c.LvsU*0 // Low Mean
lincom _cons*1 + c.LvsM*1  + c.LvsU*0 // Mid Mean
lincom _cons*1 + c.LvsM*0  + c.LvsU*1 // Upp Mean
lincom           c.LvsM*1  + c.LvsU*0 // Low vs Mid Diff
lincom           c.LvsM*0  + c.LvsU*1 // Low vs Upp Diff
lincom           c.LvsM*-1 + c.LvsU*1 // Mid vs Upp Diff
```

SAS GLM Predicting Income from 3-Category workclass

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**2**	14414.0265	7207.0132	**42.14**	**<.0001**
Error	**731**	125009.2054	**171.0112**		
Corrected Total	733	139423.2319			

R-Square	Coeff Var	Root MSE	income Mean
0.103383	75.57777	13.07713	17.30287

Mean Square Error, the residual variance, is 171.01 after including 2 slopes for workclass as a predictor (which accounted for 10.34% of the variance in income as the model R^2). The *F*-test tells us this R^2 is significantly > 0, $F(2, 731) = 42.14$, MSE = 171.01, $p < .001$.

Table of Model-Estimated Fixed Effects (normally is last)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	13.65004014	0.62628075	21.80	<.0001	12.42051668	14.87956360	**Beta0**
LvsM	8.85426742	1.00368116	8.82	<.0001	6.88382600	10.82470884	**Beta1**
LvsU	10.98470986	2.99044960	3.67	0.0003	5.11381580	16.85560393	**Beta2**

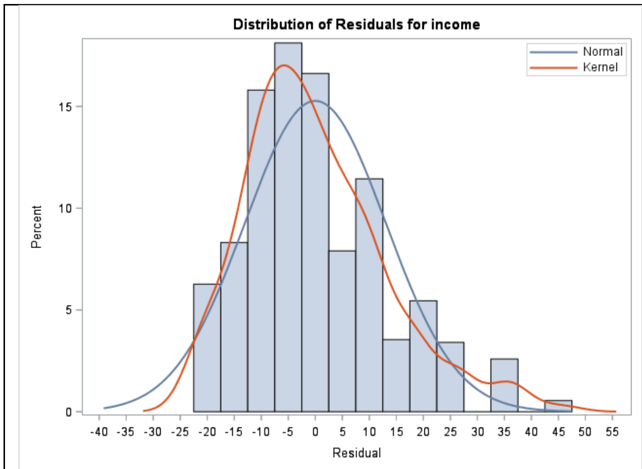
Interpret β_0 = Intercept:

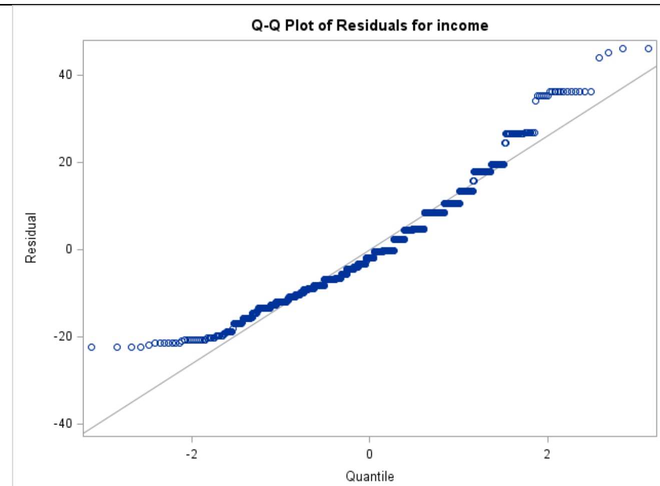
Interpret β_1 = slope of Low vs Mid:

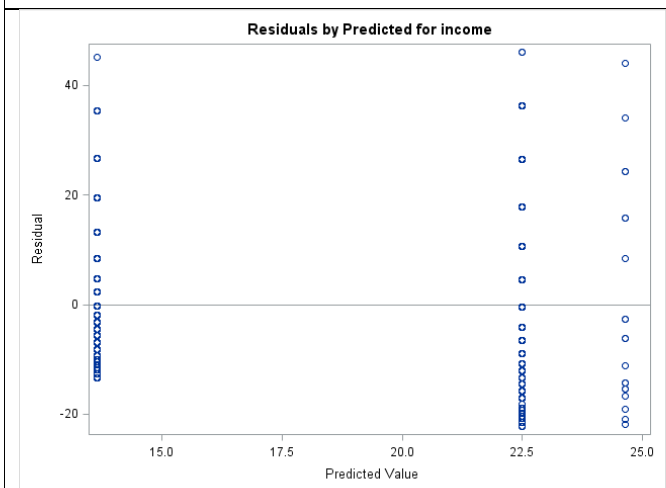
Interpret β_2 = slope of Low vs Upp:

Table of Extra Requested Linear Combinations of Model-Estimated Fixed Effects

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Low Mean	13.6500401	0.62628075	21.80	<.0001	12.4205167	14.8795636
Mid Mean	22.5043076	0.78431390	28.69	<.0001	20.9645311	24.0440840
Upp Mean	24.6347500	2.92413427	8.42	<.0001	18.8940472	30.3754528
Low vs Mid Diff	8.8542674	1.00368116	8.82	<.0001	6.8838260	10.8247088
Low vs Upp Diff	10.9847099	2.99044960	3.67	0.0003	5.1138158	16.8556039
Mid vs Upp Diff	2.1304424	3.02749229	0.70	0.4818	-3.8131743	8.0740592







Inspecting residuals for normality and constant variance (homoscedasticity)

Top: The residuals deviate from normality with fewer low cases than expected (due to income being lower-bounded at 0), as well as more high cases than expected. These trends are shown more readily in the Q-Q plot on the right, which plots the quantiles of the normal distribution on the x-axis against the quantiles of the residuals on the y-axis. Perfectly normal residuals would follow the line.

Bottom: In the plot of residuals by predicted value (which is workclass here), the residuals appear to have relatively constant variance (i.e., homoscedasticity across 3-category workclass), although perhaps there may be less variance in the lower class than in the other two classes. Because the model perfectly recreates the three means for the three categories, we do not have the potential for predictor misspecification (i.e., missing some effect of workclass).

```
TITLE "SAS PROC REG to get standardized slopes for workclass";
PROC REG DATA=work.Example4;
     MODEL income = LvsM LvsU / STB;
RUN; QUIT; TITLE;

display "STATA regress adding beta to get standardized slopes for workclass"
regress income c.LvsM c.LvsU, beta // with beta, no longer shows unstandardized CIs
```

SAS PROC REG to get standardized slopes for workclass: new information relative to GLM is in bold

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	14414	7207.01325	42.14	<.0001
Error	731	125009	171.01122		
Corrected Total	733	139423			

Root MSE	13.07713	R-Square	0.1034
Dependent Mean	17.30287	**Adj R-Sq**	**0.1009**
Coeff Var	75.57777		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	**Standardized Estimate**
Intercept	Intercept	1	13.65004	0.62628	21.80	<.0001	**0**
LvsM		1	8.85427	1.00368	8.82	<.0001	**0.31163**
LvsU		1	10.98471	2.99045	3.67	0.0003	**0.12976**

Extra SAS Syntax and Output to Compute Cohen's D Effect Sizes from Requested Estimates:

```
* Compute Cohen d effect sizes for mean differences;
DATA work.ClassEstimates; SET work.ClassEstimates;
     CohenD=(2*tvalue)/SQRT(731); * Number = denominator DF; RUN;

* Print effect sizes;
TITLE "Cohen D Effect Sizes for 3-Category Respondent Class";
PROC PRINT NOOBS DATA=work.ClassEstimates;
     VAR Parameter--CohenD; RUN;
```

Cohen D Effect Sizes for 3-Category Respondent Class

Parameter	Estimate	StdErr	tValue	Probt	LowerCL	UpperCL	CohenD
Low Mean	13.6500401	0.62628075	21.80	<.0001	12.4205167	14.8795636	1.61226
Mid Mean	22.5043076	0.78431390	28.69	<.0001	20.9645311	24.0440840	2.12250
Upp Mean	24.6347500	2.92413427	8.42	<.0001	18.8940472	30.3754528	0.62319
Low vs Mid Diff	8.8542674	1.00368116	8.82	<.0001	6.8838260	10.8247088	0.65257
Low vs Upp Diff	10.9847099	2.99044960	3.67	0.0003	5.1138158	16.8556039	0.27172
Mid vs Upp Diff	2.1304424	3.02749229	0.70	0.4818	-3.8131743	8.0740592	0.05205

<u>Example Results Section for Group Mean Differences by Working Classes:</u>

We used a general linear model (i.e., analysis of variance) to examine the extent to which annual income in thousands of dollars (M = 17.30, SD = 13.79, range = 0.25 to 68.60) could be predicted from three categories of self-reported working class membership (lower = 59.40%, middle = 37.87%, and upper = 2.72%). We created two contrasts to distinguish the three classes, in which lower-class respondents served as the reference group to be compared separately to middle-class and to upper-class respondents. We found that class membership significantly predicted annual income, $F(2, 731) = 42.14$, MSE = 171.01, $p < .001$, $R^2 = .10$. Relative to lower-class respondents, annual income was significantly higher for both middle-class respondents (Est = 8.85, SE = 1.00, d = 0.65) and upper-class respondents (Est = 10.98, SE = 2.99, d = 0.27). However, upper-class respondents did not differ significantly from middle-class respondents (Est = 2.13, SE = 3.03, d = 0.05).

Syntax and SAS Output for Age Predicting Income:

```
TITLE "SAS Descriptive Statistics for Quantitative age";
PROC MEANS NDEC=3 NOLABELS N MEAN STDDEV VAR MIN MAX DATA=work.Example4;
     VAR age;
RUN; TITLE;

display "STATA Descriptive Statistics for Quantitative age"
format age %5.3f
summarize age, format detail // detail to get variance
```

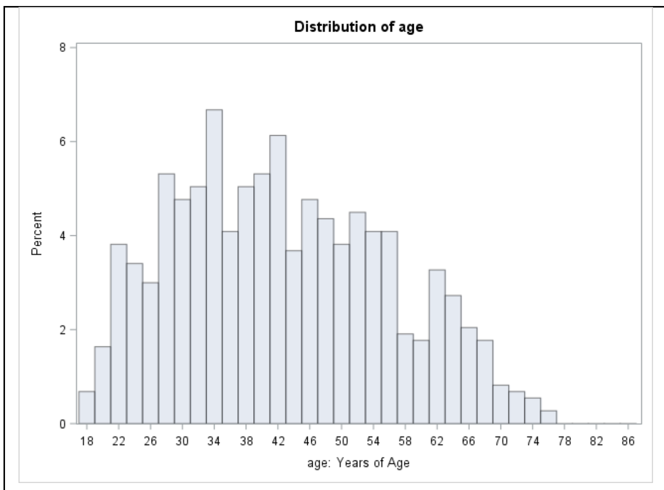
Analysis Variable : age

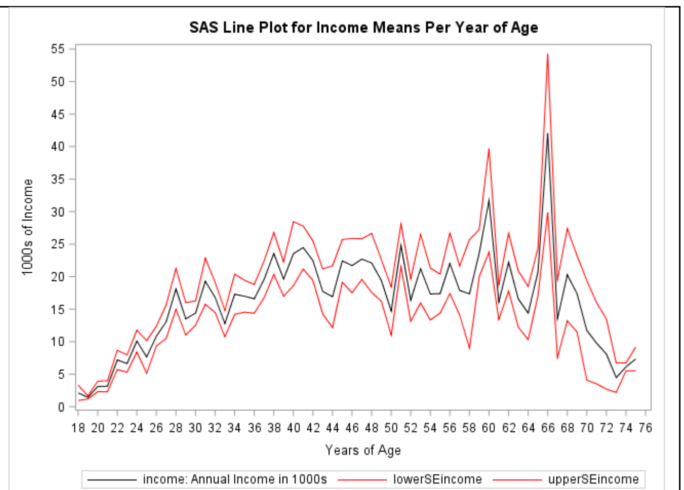
N	Mean	Std Dev	Variance	Minimum	Maximum
734	42.063	13.378	178.981	18.000	75.000

```
* Histograms to visualize quantitative age;
* NOPRINT spares the rest of the results I do not want right now;
TITLE "SAS Histogram of Quantitative age";
PROC UNIVARIATE NOPRINT DATA=work.Example4;
     VAR age;
     HISTOGRAM age / MIDPOINTS=18 TO 86 BY 2;
RUN; QUIT; TITLE;

display "STATA Histogram of Quantitative age"
histogram age, percent discrete width(2) start(18)
```

Left: Histogram for age; looks like we can treat this as a typical quantitative variable (i.e., no notable piles of same value)

Right: plot of means for income (as outcome) per year of age (predictor); see code online






```
* SAS code to make new age variable centered at 18 (minimum in sample);
DATA work.Example4; SET work.Example4;
     age18=age-18;
     LABEL age18= "age18: Age (0=18 years)";
RUN;

// STATA code to make new age variable centered at 18 (minimum in sample)
   gen age18=age-18
   label variable age18 "age18: Age (0=18 years)"
```

First Testing a <u>Linear</u> Effect of Age (0=18): $Income_i = \beta_0 + \beta_1(Age_i - 18) + e_i$

```
TITLE "SAS GLM Predicting Income from Linear Centered Age (0=18)";
PROC GLM DATA=work.Example4 NAMELEN=100; * PLOTS(UNPACK)=DIAGNOSTICS;
     MODEL income = age18 / SOLUTION ALPHA=.05 CLPARM;
     * Ask for predicted income for example ages;
```

```
     ESTIMATE "Pred Income Age 30 (age18=12)" intercept 1 age18 12;
     ESTIMATE "Pred Income Age 50 (age18=32)" intercept 1 age18 32;
     ESTIMATE "Pred Income Age 70 (age18=52)" intercept 1 age18 52;
RUN; QUIT; TITLE;

display "STATA GLM Predicting Income from Linear Centered Age (0=18)"
regress income c.age18, level(95)
// Ask for predicted income for example ages
   lincom _cons*1 + c.age18*12  // Pred Income Age 30 (age18=12)
   lincom _cons*1 + c.age18*32  // Pred Income Age 50 (age18=32)
   lincom _cons*1 + c.age18*52  // Pred Income Age 70 (age18=52)
```

SAS GLM Predicting Income from Linear Centered Age (0=18)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	5580.7424	5580.7424	30.52	<.0001
Error	732	133842.4895	182.8449		
Corrected Total	733	139423.2319			

R-Square	Coeff Var	Root MSE	income Mean
0.040027	78.14896	13.52202	17.30287

Mean Square Error, the residual variance, is 182.84 after including a linear effect of age (which accounted for 4.00% of the variance in income as the model R^2). The *F*-test tells us this R^2 is significantly > 0, $F(1, 732) = 30.52$, MSE = 182.84, $p < .001$.

Table of Model-Estimated Fixed Effects (normally is last)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	12.33998883	1.02765825	12.01	<.0001	10.32247980	14.35749786
age18	0.20624834	0.03733240	5.52	<.0001	0.13295699	0.27953969

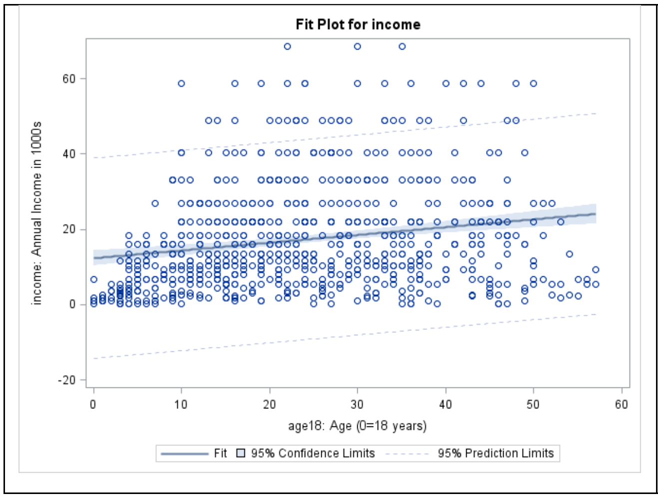
Interpret β_0 = Intercept:

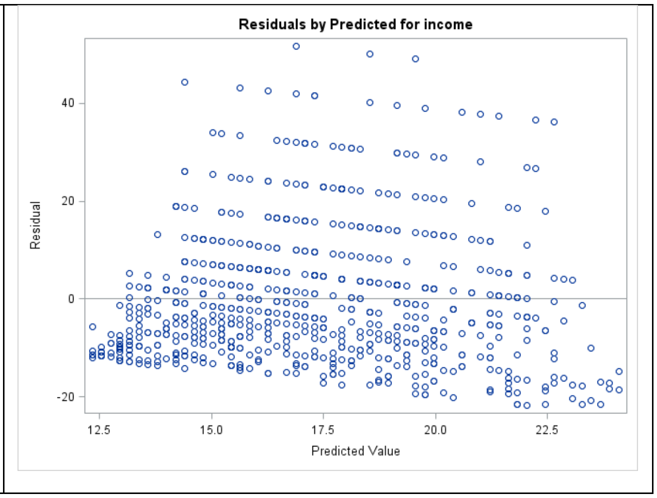
Interpret β_1 = slope of age18:

Table of Extra Requested Linear Combinations of Model-Estimated Fixed Effects

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Pred Income Age 30 (age18=12)	14.8149689	0.67223750	22.04	<.0001	13.4952255	16.1347124
Pred Income Age 50 (age18=32)	18.9399357	0.58044193	32.63	<.0001	17.8004063	20.0794652
Pred Income Age 70 (age18=52)	23.0649026	1.15623917	19.95	<.0001	20.7949622	25.3348429

Left: model-predicted regression line through scatterplot for the linear effect of age
Right: model residuals by age—these should be flat and even (→ constant variance), but there is a negative trend visible indicating misspecification of the effect of age (i.e., there is an effect of age beyond just linear that should be included)





Now Adding a Quadratic Effect of Age (0=18):

$$Income_i = \beta_0 + \beta_1(Age_i - 18) + \beta_2(Age_i - 18)^2 + e_i$$

```
TITLE "SAS GLM Predicting Income from Quadratic Centered Age (0=18)";
PROC GLM DATA=work.Example4 NAMELEN=100; * PLOTS(UNPACK)=DIAGNOSTICS;
     * Asterisk creates multiplied predictor variable;
     MODEL income = age18 age18*age18 / SOLUTION ALPHA=.05 CLPARM;
     * Save predicted income and SE to new dataset to make pictures;
     OUTPUT OUT=work.PredIncomebyAge PREDICTED=YhatAge STDP=SEyhatAge;
     * Ask for predicted income for example ages;
     ESTIMATE "Pred Income Age 30 (age18=12)" intercept 1 age18 12 age18*age18 144;
     ESTIMATE "Pred Income Age 50 (age18=32)" intercept 1 age18 32 age18*age18 1024;
     ESTIMATE "Pred Income Age 70 (age18=52)" intercept 1 age18 52 age18*age18 2704;
     * Linear age slope changes by 2*quadratic coefficient, so multiply age*2;
     ESTIMATE "Pred Linear Age Slope Age 30 (age18=12)"   age18 1  age18*age18 24;
     ESTIMATE "Pred Linear Age Slope Age 50 (age18=32)"   age18 1  age18*age18 64;
     ESTIMATE "Pred Linear Age Slope Age 70 (age18=52)"   age18 1  age18*age18 104;
RUN; QUIT; TITLE;

display as result "STATA GLM Predicting Income from Quadratic Centered Age (0=18)"
regress income c.age18 c.age18#c.age18, level(95) // Hashtag (pound) multiplies predictors
// Ask for predicted income for example ages
   lincom _cons*1 + c.age18*12 + c.age18#c.age18*144  // Pred Income Age 30 (age18=12)
   lincom _cons*1 + c.age18*32 + c.age18#c.age18*1024 // Pred Income Age 50 (age18=32)
   lincom _cons*1 + c.age18*52 + c.age18#c.age18*2704 // Pred Income Age 70 (age18=52)
// Linear age slope changes by 2*quadratic coefficient, so multiply age*2
   lincom  c.age18*1 + c.age18#c.age18*24  // Pred Linear Age Slope at Age 30 (age18=12)
   lincom  c.age18*1 + c.age18#c.age18*64  // Pred Linear Age Slope at Age 50 (age18=32)
   lincom  c.age18*1 + c.age18#c.age18*104 // Pred Linear Age Slope at Age 70 (age18=52)
```

SAS GLM Predicting Income from Quadratic Centered Age (0=18)

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**2**	15885.4618	7942.7309	**47.00**	**<.0001**
Error	**731**	123537.7701	**168.9983**		
Corrected Total	733	139423.2319			

R-Square	Coeff Var	Root MSE	income Mean
0.113937	75.13165	12.99994	17.30287

Mean Square Error, the residual variance, is now 169.00 from the two effects of age (which accounted for 11.39% of the variance in income as the model R^2). The F-test says this R^2 is significantly > 0, $F(2, 731) = 47.00$, MSE = 169.00, $p < .001$.

Table of Model-Estimated Fixed Effects (normally is last)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	2.676597431	1.58352919	1.69	0.0914	-0.432210062	5.785404923	**Beta0**
age18	1.223080607	0.13507406	9.05	<.0001	0.957901252	1.488259961	**Beta1**
age18*age18	-0.019537211	0.00250199	-7.81	<.0001	-0.024449155	-0.014625267	**Beta2**

Interpret β_0 = Intercept:

Interpret β_1 = slope of age18:

Interpret β_2 = slope of age18^2:

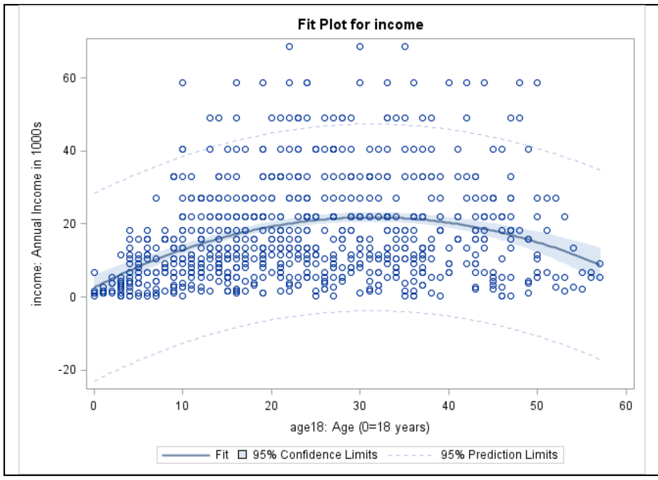
Interpret R^2 two different ways:

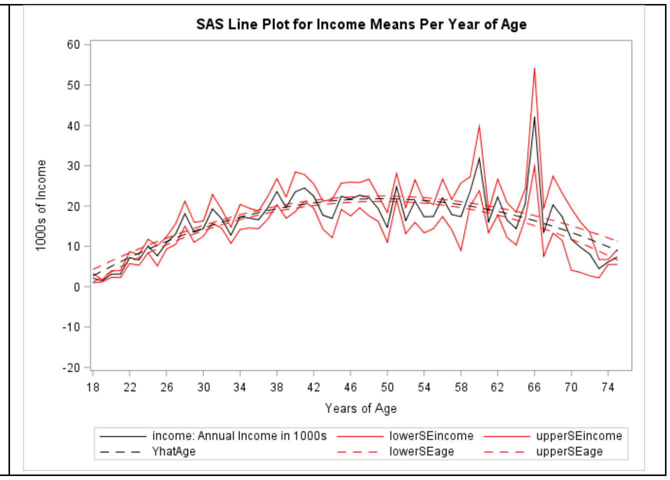
The R^2 went from .040 to .114, an increase of .074. Do we know if the R^2 increased significantly relative to the linear age model?

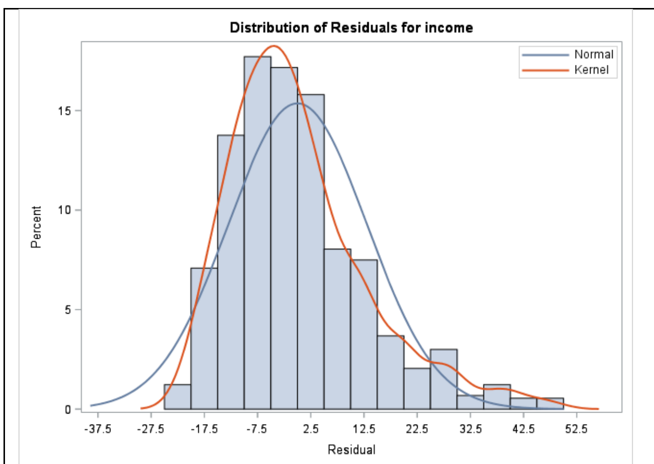
Table of Extra Requested Linear Combinations of Model-Estimated Fixed Effects

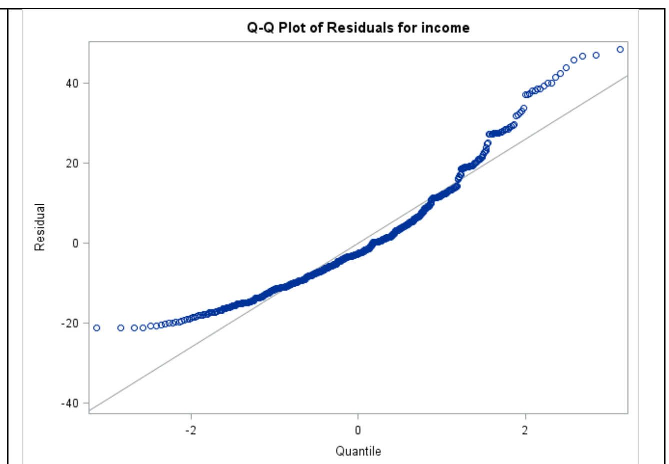
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Pred Income Age 30 (age18=12)	14.5402064	0.64723977	22.46	<.0001	13.2695359	15.8108769
Pred Income Age 50 (age18=32)	21.8090730	0.66813438	32.64	<.0001	20.4973819	23.1207641
Pred Income Age 70 (age18=52)	13.4481710	1.65902182	8.11	<.0001	10.1911553	16.7051867
Pred Linear Age Slope Age 30 (age18=12)	0.7541875	0.07881678	9.57	<.0001	0.5994533	0.9089218
Pred Linear Age Slope Age 50 (age18=32)	-0.0273009	0.04671950	-0.58	0.5592	-0.1190213	0.0644195
Pred Linear Age Slope Age 70 (age18=52)	-0.8087893	0.13485251	-6.00	<.0001	-1.0735337	-0.5440449

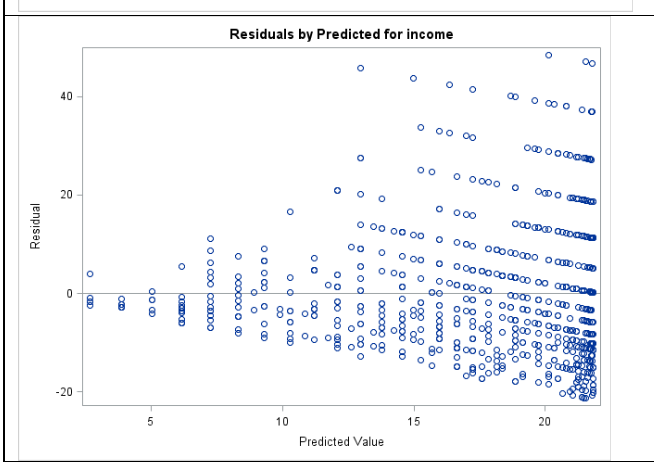
Left: model-predicted regression line through scatterplot (provided automatically)
Right: model-predicted regression line through means for age (see extra code online)











Inspecting residuals for normality and constant variance (homoscedasticity)

Top: The residuals still deviate from normality with fewer low cases and more high cases than expected.

Bottom: In the plot of residuals by predicted value, no trend is visible (a flat line would fit), indicating we have reasonably specified the effect of age. But the residuals appear to have much greater variability in richer persons, indicating potential heterogeneity of variance. It would be better to fit a model that allows the residual variance to differ quadratically by age (i.e., using PROC MIXED).

We forgo requesting standardized slopes for this model given the ambiguity of how to interpret them for models with interactions… R^2 is a sufficiently useful effect size to describe the overall effect (trend) of age here.

Example Results Section for the Linear and Quadratic Effects of Age:

We used a general linear model (i.e., linear regression) to examine the extent to which annual income in thousands of dollars (M = 17.30, SD = 13.79, range = 0.25 to 68.60) could be predicted from years of age (M = 42.06, SD = 13.38, range = 18 to 75). We first examined the means of income by age to identify plausible types of nonlinear associations. Given the apparent curvilinear trend (in which age appeared positively associated with income until middle age, upon which it appeared negatively associated instead), we fit a model including linear and quadratic slopes for age (in which age was centered such that 0 = 18 years, the minimum age in the sample). The quadratic age model captured a significant amount of variance in annual income, $F(2, 731) = 47.00$, MSE = 169.00, $p < .001$, $R^2 = .114$. The quadratic age model was also a significant improvement over a linear age model, as indicated by the significant slope for the quadratic effect of age. The model fixed effects can be interpreted as follows. The fixed intercept indicated that at age 18, annual income was predicted to be 2.676 thousand dollars (SE = 1.584) and was expected to be significantly greater by 1.223 thousand dollars per year of age (i.e., the instantaneous linear slope for age at age 18; SE = 0.135, $p < .001$). The linear age slope at age 18 was predicted to become significantly more negative per year of age by twice the quadratic coefficient of −0.020 (SE = 0.002, $p < .001$). As given by the quantity (−1*linear slope) / (2*quadratic slope) + 18, the age of maximum predicted personal income was 48.575 (i.e., the age at which the linear age slope = 0). For example, the linear effect of age as evaluated at age 30 was significantly positive (Est = 0.754, SE = 0.079), the linear effect of age as evaluated at age 50 was nonsignificantly negative (Est = −0.027, SE = 0.047), and the linear effect of age as evaluated at age 70 was significantly negative (Est = −0.809, SE = 0.135).

Syntax and SAS Output with Education Predicting Income:

```
TITLE "SAS Descriptive Statistics for Quantitative Variable education";
PROC MEANS NDEC=3 NOLABELS N MEAN STDDEV VAR MIN MAX DATA=work.Example4;
     VAR educ;
RUN; TITLE;

display "STATA Descriptive Statistics for Quantitative education"
format educ %5.3f
summarize educ, format detail // detail to get variance
```

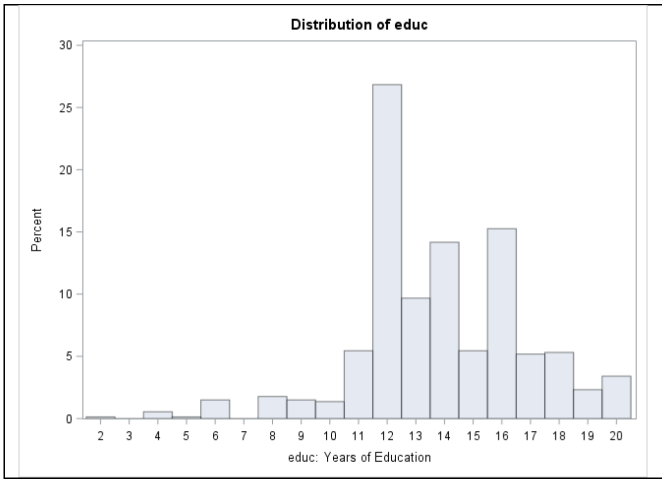
Analysis Variable : educ

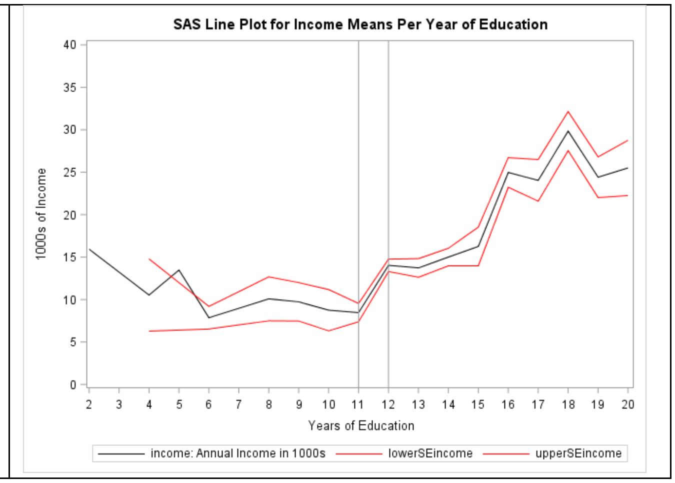
N	Mean	Std Dev	Variance	Minimum	Maximum
734	13.812	2.909	8.464	2.000	20.000

```
* Histograms to visualize quantitative variables;
* NOPRINT spares the rest of the results I do not want right now;
TITLE "SAS Histograms of Quantitative Variable education";
PROC UNIVARIATE NOPRINT DATA=work.Example4;
     VAR educ;
     HISTOGRAM educ / MIDPOINTS=2 TO 20 BY 1;
RUN; QUIT; TITLE;

display "STATA Histogram of Quantitative education"
histogram educ, percent discrete width(1) start(2)
```

Left: Histogram for education; looks like 12 is somewhat of a break point

Right: plot of means for income (as outcome) per year of education (predictor); see code online—points to "sections" of different slopes for ed






```
* SAS code to make 3 new variables for sections of education;
DATA work.Example4; SET work.Example4;
     lessHS=.; gradHS=.; overHS=.; * Make three new empty variables;
     * Replace for educ less than 12;
     IF educ LT 12 THEN DO; lessHS=educ-11; gradHS=0; overHS=0;        END;
     * Replace for educ greater or equal to 12;
     IF educ GE 12 THEN DO; lessHS=0;       gradHS=1; overHS=educ-12; END;
     LABEL lessHS= "lessHS: Slope for Years Ed Less Than High School"
           gradHS= "gradHS: Bump for Graduating High School"
           overHS= "overHS: Slope for Years Ed After High School";
RUN;
```

Years Educ (x)	lessHS: Slope if x <12	gradHS: HS Grad? (0=no, 1=yes)	overHS: Slope if x >12
9	-2	0	0
10	-1	0	0
11 (int)	0	0	0
12	0	1	0
13	0	1	1
14	0	1	2
15	0	1	3
16	0	1	4
17	0	1	5
18	0	1	6

```
// STATA code to make 3 new variables for sections of education
   gen lessHS=. // Make 3 new empty variables
   gen gradHS=.
   gen overHS=.
// Replace for educ less than 12
   replace lessHS=educ-11 if educ <  12
   replace gradHS=0       if educ <  12
   replace overHS=0       if educ <  12
// Replace for educ greater or equal to 12
   replace lessHS=0       if educ >= 12
   replace gradHS=1       if educ >= 12
   replace overHS=educ-12 if educ >= 12
// Label variables
   label variable lessHS "lessHS: Slope for Years Ed Less Than High School"
   label variable gradHS "gradHS: Acute Bump for Graduating High School"
   label variable overHS "overHS: Slope for Years Ed After High School"
```

Piecewise Linear Effects of Education:

$$Income_i = \beta_0 + \beta_1(LessHS_i) + \beta_2(GradHS_i) + \beta_3(OverHS_i) + e_i$$

```
TITLE "SAS GLM Predicting Income from Piecewise Education";
PROC GLM DATA=work.Example4 NAMELEN=100; * PLOTS(UNPACK)=DIAGNOSTICS;
     MODEL income = lessHS gradHS overHS / SOLUTION ALPHA=.05 CLPARM;
     * Example of how to compare slopes;
     ESTIMATE "Diff in ed slope:  2-11 vs 11-12" lessHS -1 gradHS 1;
     ESTIMATE "Diff in ed slope: 11-12 vs 12-20" gradHS -1 overHS 1;
     * Save predicted income and SE to new dataset to make pictures;
     OUTPUT OUT=work.PredIncomebyEduc PREDICTED=YhatEduc STDP=SEyhatEduc;
RUN; QUIT; TITLE;

display "STATA GLM Predicting Income from Piecewise Education"
regress income c.lessHS c.gradHS c.overHS, level(95)
```

```
// Example of how to compare slopes
   lincom c.lessHS*-1 + c.gradHS*1 // Diff in ed slope:  2-11 vs 11-12
   lincom c.gradHS*-1 + c.overHS*1 // Diff in ed slope: 11-12 vs 12-20
```

SAS GLM Predicting Income from Piecewise Education

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**3**	22906.5605	7635.5202	**47.84**	**<.0001**
Error	**730**	116516.6714	**159.6119**		
Corrected Total	733	139423.2319			

R-Square	Coeff Var	Root MSE	income Mean
0.164295	73.01538	12.63376	17.30287

Mean Square Error, the residual variance, is 159.61 given the piecewise education slopes (which accounted for 16.43% of the variance in income as the model R^2). The *F*-test says this R^2 is significantly > 0, $F(3, 730) = 47.84$, MSE = 159.61, $p < .001$.

Table of Model-Estimated Fixed Effects (normally is last)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	8.534867248	1.72935077	4.94	<.0001	5.139773001	11.929961495	**Beta0**
lessHS	-0.268784499	0.59880153	-0.45	0.6537	-1.444363022	0.906794023	**Beta1**
gradHS	4.684746178	1.87568395	2.50	0.0127	1.002367857	8.367124499	**Beta2**
overHS	2.124528973	0.21372442	9.94	<.0001	1.704941139	2.544116806	**Beta3**

Interpret $\boldsymbol{\beta_0}$ = Intercept:

Interpret $\boldsymbol{\beta_1}$ = slope of lessHS:

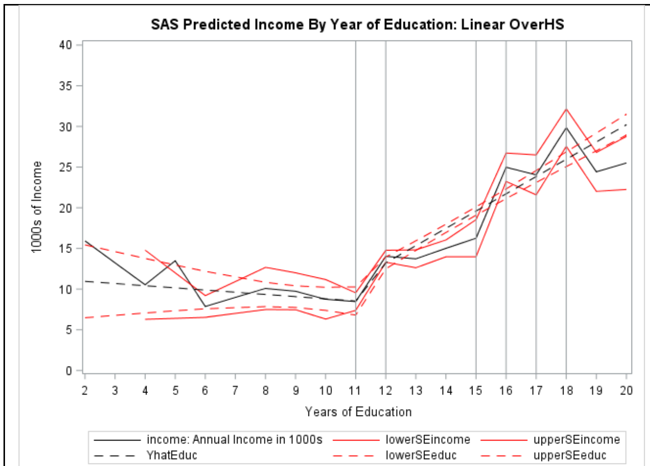
Interpret $\boldsymbol{\beta_2}$ = slope of gradHS:

Interpret $\boldsymbol{\beta_3}$ = slope of overHS:

Table of Extra Requested Linear Combinations of Model-Estimated Fixed Effects

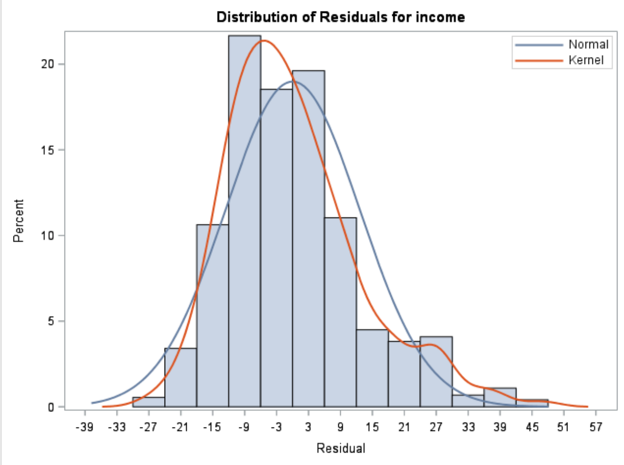
Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Diff in ed slope: 2-11 vs 11-12	4.95353068	2.28222698	2.17	0.0303	0.47301937	9.43404199
Diff in ed slope: 11-12 vs 12-20	-2.56021721	1.94673385	-1.32	0.1889	-6.38208203	1.26164762

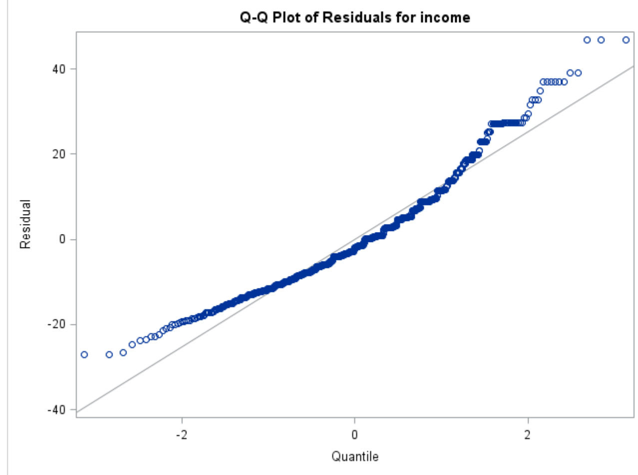
Comparisons of Slopes Above: The slope for gradHS is significantly more positive than the slope for lessHS (indicating that they should not be constrained to be the same). The slope for overHS is nonsignificantly less positive than the slope for gradHS (indicating that they *could* be constrained to be the same). However, it's important to note that the slope for overHS—implying a linear effect of each additional year of education—does not appear to fit the means well. So efforts to refine the model should focus on better capturing differences by education after 12 years first…

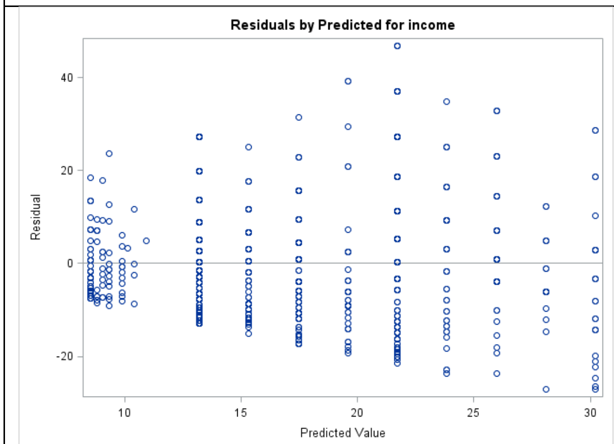


Left: model-predicted regression line through means for education (see extra code online)

As shown by the misfit of the data to the model (dashed line), it looks like the effect of education after 12 years should have additional piecewise slopes (i.e., 12–15, 15–17, 17–18, 18–20)… if you are feeling brave, give it a try and let me know what happens!







Inspecting residuals for normality and constant variance (homoscedasticity)

Top: The residuals still deviate from normality with fewer low cases and more high cases than expected.

Bottom: In the plot of residuals by predicted value, a negative trend is apparent, indicating we may have mis-specified the effect of education. The residuals also appear to have much greater variability in more educated persons, indicating potential heterogeneity of variance. It would be better to fit a model that allows the residual variance to increase by education (i.e., using PROC MIXED).

```
TITLE "SAS PROC REG to get standardized slopes for education";
PROC REG DATA=work.Example4;
     MODEL income = lessHS gradHS overHS / STB;
RUN; QUIT; TITLE;
```

```
display "STATA regress adding beta to get standardized slopes for education"
regress income c.lessHS c.gradHS c.overHS, beta // with beta, no longer shows CIs
```

SAS PROC REG to get standardized slopes for education: new information relative to GLM is in bold

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	22907	7635.52017	47.84	<.0001
Error	730	116517	159.61188		
Corrected Total	733	139423			

Root MSE	12.63376	R-Square	0.1643
Dependent Mean	17.30287	**Adj R-Sq**	**0.1609**
Coeff Var	73.01538		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	**Standardized Estimate**
Intercept	1	8.53487	1.72935	4.94	<.0001	0
lessHS	1	-0.26878	0.59880	-0.45	0.6537	**-0.01932**
gradHS	1	4.68475	1.87568	2.50	0.0127	**0.11202**
overHS	1	2.12453	0.21372	9.94	<.0001	**0.35903**

Example Results Section for Piecewise Effect of Education:

We used a general linear model (i.e., linear regression) to examine the extent to which annual income in thousands of dollars (*M* = 17.30, *SD* = 13.79, range = 0.25 to 68.60) could be predicted from years of education (*M* = 13.81, *SD* = 2.91, range = 2 to 20). We first examined the means of income by education to identify plausible types of nonlinear associations. The effect of education predicting annual income appeared to differ across regions of education, suggesting a piecewise trend with the distinct region slopes to be captured by linear splines. Specifically, we fit one linear slope for the effect of education from 2 to 11 years, a second linear slope of education from 11 to 12 years, and a third linear slope of education from 12 to 20 years. The model including these three education slopes captured a significant amount of variance in annual income, $F(3, 730) = 47.84$, MSE = 159.61, $p < .001$, $R^2 = .164$. The model fixed effects can be interpreted as follows. Annual income was expected to be nonsignificantly lower by 0.27 thousand dollars per year of education from 2 to 11 years (*SE* = 0.60, $p = .654$, $Est_{std} = -.019$), resulting in predicted annual income of 8.53 thousand dollars (SE = 1.73) at 11 years of education (i.e., as given by the fixed intercept). Annual income was then expected to be significantly higher by 4.68 thousand dollars (SE =1.88, $p = .013$, $Est_{std} = .112$) for those achieving a high school degree (i.e., a significant difference between 11 and 12 years of education). Although annual income was expected to be significantly higher by 2.12 thousand dollars (SE = 0.21, $p < .001$, $Est_{std} = 0.359$) per year of additional education past 12 years, examining a plot of the observed versus predicted means for annual income at each year of education suggested a linear slope was not sufficient in capturing the observed differences in income from 12 to 20 years of education. We recommend considering in future research the use of additional piecewise slopes corresponding to distinct levels of higher education (e.g., bachelors, masters, or doctoral college degrees).

Syntax and SAS Output with 5-Category Ordinal Happiness Predicting Income:

```
TITLE "SAS Descriptive Statistics for Categorical happy";
PROC FREQ DATA=work.Example4;
     TABLE happy;
RUN; TITLE;

display "STATA Descriptive Statistics for Categorical happy"
tabulate happy
```

happy: 5-Category Happy Rating

happy	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1.Unhappy	26	3.54	26	3.54
2.Neither	39	5.31	65	8.86
3.Fairly Happy	256	34.88	321	43.73
4.Very Happy	327	44.55	648	88.28
5.Completely Happy	86	11.72	734	100.00

```
* SAS code to make a single centered predictor for happy;
DATA work.Example4; SET work.Example4;
     happy1=happy-1;
     LABEL happy1= "happy1: Happy Category (0=1)";
RUN;

// STATA code to make a single centered predictor for happy
   gen happy1=happy-1
   label variable happy1 "happy1: Happy Category (0=1)"
```

First Testing a <u>Linear</u> Effect of Happy (0=1): $Income_i = \beta_0 + \beta_1(Happy_i - 1) + e_i$

```
TITLE "SAS GLM Predicting Income from Linear Centered Happy (0=1)";
PROC GLM DATA=work.Example4 NAMELEN=100; * PLOTS(UNPACK)=DIAGNOSTICS;
     MODEL income = happy1 / SOLUTION ALPHA=.05 CLPARM;
     * Save predicted income and SE to new dataset to make pictures;
     OUTPUT OUT=work.PredIncomebyHappy1 PREDICTED=Yhat1Happy STDP=SEyhat1Happy;
RUN; QUIT; TITLE;
```

```
display "STATA GLM Predicting Income from Linear Centered Happy (0=1)"
regress income c.happyc1, level(95)
```

```
SAS GLM Predicting Income from Linear Centered Happy (0=1)
                                          Sum of
Source                      DF           Squares      Mean Square    F Value    Pr > F
Model                        1          320.3981         320.3981       1.69    0.1945
Error                      732       139102.8338         190.0312
Corrected Total            733       139423.2319

R-Square     Coeff Var      Root MSE    income Mean
0.002298      79.66988      13.78518       17.30287
```

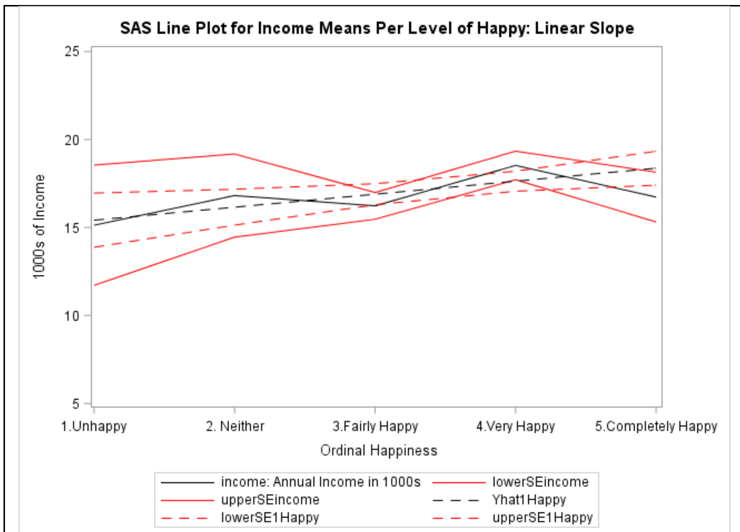
Mean Square Error, the residual variance, is 190.03 after a linear effect of happy (which accounted for 0.23% of the variance in income as the model R^2). The *F*-test tells us this R^2 is **not** significantly > 0, $F(1, 732) = 1.69$, MSE = 190.03, $p = .195$.

Table of Model-Estimated Fixed Effects (normally is last)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	15.41494808	1.54042165	10.01	<.0001	12.39077678	18.43911937	Beta0
happy1	0.73866637	0.56887362	1.30	0.1945	-0.37815205	1.85548479	Beta1

Interpret β_0 = Intercept:

Interpret β_1 = slope of happy1:



Left: model-predicted regression line through means for age (see extra code online)

In addition to not really making sense (i.e., these values are ordinal, not interval, so they aren't really numbers), a single linear slope predicting the same difference between each pair of happiness categories doesn't seem to fit the pattern of means.

So let's fit a piecewise slopes model in which the slopes capture each shift between adjacent categories…

```
* SAS code to make 4 new variables for adjacent values of happy;
DATA work.Example4; SET work.Example4;
     h1v2=.; h2v3=.; h3v4=.; h4v5=.; * Make 4 new empty variables;
     IF happy=1 THEN DO; h1v2=0; h2v3=0; h3v4=0; h4v5=0; END;
     IF happy=2 THEN DO; h1v2=1; h2v3=0; h3v4=0; h4v5=0; END;
     IF happy=3 THEN DO; h1v2=1; h2v3=1; h3v4=0; h4v5=0; END;
     IF happy=4 THEN DO; h1v2=1; h2v3=1; h3v4=1; h4v5=0; END;
     IF happy=5 THEN DO; h1v2=1; h2v3=1; h3v4=1; h4v5=1; END;
     LABEL h1v2="Slope from Happy 1 to 2"
           h2v3="Slope from Happy 2 to 3"
           h3v4="Slope from Happy 3 to 4"
           h4v5="Slope from Happy 4 to 5";
RUN;

// STATA code to make 4 new variables for adjacent values of happy
// Make 4 new empty variables
   gen h1v2=.
   gen h2v3=.
   gen h3v4=.
   gen h4v5=.
```

```
// Replace with 0s
   replace h1v2=0 if happy < 2
   replace h2v3=0 if happy < 3
   replace h3v4=0 if happy < 4
   replace h4v5=0 if happy < 5
// Replace with 1s
   replace h1v2=1 if happy >= 2
   replace h2v3=1 if happy >= 3
   replace h3v4=1 if happy >= 4
   replace h4v5=1 if happy == 5
// Label variables
   label variable h1v2 "Slope from Happy 1 to 2"
   label variable h2v3 "Slope from Happy 2 to 3"
   label variable h3v4 "Slope from Happy 3 to 4"
   label variable h4v5 "Slope from Happy 4 to 5"
```

Piecewise Adjacent Slopes of Happy:

$$\boldsymbol{Income_i} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}(\boldsymbol{h1v2_i}) + \boldsymbol{\beta_2}(\boldsymbol{h2v3_i}) + \boldsymbol{\beta_3}(\boldsymbol{h3v4_i}) + \boldsymbol{\beta_3}(\boldsymbol{h4v5_i}) + \boldsymbol{e_i}$$

```
TITLE "SAS GLM Predicting Income from Piecewise Adjacent Slopes for Happy";
PROC GLM DATA=work.Example4 NAMELEN=100; * PLOTS(UNPACK)=DIAGNOSTICS;
     MODEL income = h1v2 h2v3 h3v4 h4v5 / SOLUTION ALPHA=.05 CLPARM;
     * Example of how to compare slopes;
     ESTIMATE "Diff in Slope 1-2 vs 2-3" h1v2 -1 h2v3 1;
     ESTIMATE "Diff in Slope 2-3 vs 3-4" h2v3 -1 h3v4 1;
     ESTIMATE "Diff in Slope 3-4 vs 4-5" h3v4 -1 h4v5 1;
RUN; QUIT; TITLE;
```

```
display "STATA GLM Predicting Income from Piecewise Adjacent Slopes for Happy"
regress income c.h1v2 c.h2v3 c.h3v4 c.h4v5, level(95)
// Example of how to compare slopes
   lincom c.h1v2*-1 + c.h2v3*1 // Diff in Slope 1-2 vs Slope 2-3
   lincom c.h2v3*-1 + c.h3v4*1 // Diff in Slope 2-3 vs Slope 3-4
   lincom c.h3v4*-1 + c.h4v5*1 // Diff in Slope 3-4 vs Slope 4-5
```

SAS GLM Predicting Income from Piecewise Adjacent Slopes for Happy

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	**4**	946.3348	236.5837	**1.25**	**0.2902**
Error	**729**	138476.8971	**189.9546**		
Corrected Total	733	139423.2319			

R-Square	Coeff Var	Root MSE	income Mean
0.006787	79.65383	13.78240	17.30287

Mean Square Error, the residual variance, is 189.95 after adding the 4 slopes of happy (which accounted for 0.68% of the variance in income as the model R^2). The *F*-test tells us this R^2 is **not** significantly > 0, $F(4, 729) = 1.25$, MSE = 189.95, $p = .290$.

Table of Model-Estimated Fixed Effects (normally is last)

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits		
Intercept	15.12875000	2.70295132	5.60	<.0001	9.82225260	20.43524740	**Beta0**
h1v2	1.68516026	3.48949515	0.48	0.6293	-5.16549843	8.53581894	**Beta1**
h2v3	-0.58648838	2.36910124	-0.25	0.8045	-5.23756348	4.06458671	**Beta2**
h3v4	2.29929831	1.15017869	2.00	0.0460	0.04124054	4.55735608	**Beta3**
h4v5	-1.79692367	1.67023208	-1.08	0.2823	-5.07596246	1.48211511	**Beta4**

The fixed intercept gives the mean for x=1, and each slope gives the difference to the next category.

Table of Extra Requested Linear Combinations of Model-Estimated Fixed Effects

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Diff in Slope 3-4 vs 4-5	-2.27164864	5.24694941	-0.43	0.6652	-12.57258275	8.02928547
Diff in Slope 4-5 vs 5-6	2.88578669	2.90164986	0.99	0.3203	-2.81080036	8.58237373
Diff in Slope 5-6 vs 6-7	-4.09622198	2.29660358	-1.78	0.0749	-8.60496798	0.41252402

Comparisons of Slopes Above: No pairwise differences between slopes are significant, which means we would not lose anything predictive informative by constraining the slopes to be equal.

```
TITLE "SAS PROC REG to get standardized slopes for happy";
PROC REG DATA=work.Example4;
     MODEL income = h1v2 h2v3 h3v4 h4v5 / STB;
RUN; QUIT; TITLE;

display "STATA regress adding beta to get standardized slopes for happy"
regress income c.h1v2 c.h2v3 c.h3v4 c.h4v5, beta // with beta, no longer shows CIs
```

SAS PROC REG to get standardized slopes for happy: new information relative to GLM is in bold

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	946.33479	236.58370	1.25	0.2902
Error	729	138477	189.95459		
Corrected Total	733	139423			

Root MSE	13.78240	R-Square	0.0068
Dependent Mean	17.30287	**Adj R-Sq**	**0.0013**
Coeff Var	79.65383		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	**Standardized Estimate**
Intercept	1	15.12875	2.70295	5.60	<.0001	0
h1v2	1	1.68516	3.48950	0.48	0.6293	0.02260
h2v3	1	-0.58649	2.36910	-0.25	0.8045	-0.01209
h3v4	1	2.29930	1.15018	2.00	0.0460	0.08276
h4v5	1	-1.79692	1.67023	-1.08	0.2823	-0.04193

Example Results Section for the Linear and Piecewise Effects of Happy:

We used a general linear model (i.e., linear regression) to examine the extent to which annual income in thousands of dollars (M = 17.30, SD = 13.79, range = 0.25 to 68.60) could be predicted from ordinal happiness (unhappy = 3.54%, neither = 5.31%, fairly happy = 34.88%, very happy = 44.55%, completely happy = 11.72%). In first examining a linear effect of happiness (centered at unhappy = 0), the model fixed effects indicated that annual income was predicted to be 15.42 thousand dollars (SE = 1.54) for unhappy respondents (i.e., as given by the fixed intercept), and that annual income was predicted to be nonsignificantly greater by 0.74 thousand dollars (SE = 0.57, p = .195, R^2 = .002) per additional ordinal level of happiness.

However, given that a linear slope for happiness assumes interval differences with respect to predicted income, we tested this assumption by specifying a piecewise slopes model by which to estimate all adjacent differences in predicted annual income by ordinal level of happiness. The revised model—predicting four adjacent differences across the five levels of happiness—did not capture a significant amount of variance in annual income, $F(4, 729)$ = 1.25, MSE = 189.95, p = .290, R^2 = .007. The model fixed effects indicated that annual income was 15.13 thousand dollars (SE = 2.70) for unhappy respondents (i.e., as given by the fixed intercept). Annual income was nonsignificantly higher by 1.69 thousand dollars (SE = 3.49, p = .629) for neither than unhappy respondents, nonsignificantly lower by 0.59 thousand dollars (SE = 2.37, p = .804) for fairly happy than neither respondents, significantly higher by 2.30 thousand dollars (SE = 1.15, p = .046) for very happy than fairly happy respondents, and nonsignificantly lower by 1.80 thousand dollars (SE = 1.67, p = .282) for completely happy than very happy respondents. None of the differences between these adjacent differences were significant (as given by linear combinations of the model fixed effects, requested separately). Thus, there is little evidence that annual income can be predicted by self-rated happiness, whether treated as interval (through a linear slope) or treated as ordinal (through piecewise slopes).