MDA ASSIGNMENT 3: REGRESSION MODELS WITH NON-LINEARITIES

```
GET
  FILE='U:\)Research\ISSP20072008\issp_2007_2008nl_def.sav'.

desc age gender z06a1 z34a.

freq z06a1 z34a.

recode z34a
(1=0)(2=400)(3=800)(4=1150)(5=1450)(6=1750)(7=2150)(8=2700)(9=3500)(10=4500)
(11=5500)
  into pinc.
recode age (25 thru 34=30)(35 thru 44=40)(45 thru 54=50)(55 thru 64=60) into
agecat.
recode z06a1 (1 2 3 6=0)(4 5 7=1)(8 9=2) into educat.
add value labels educat (0) Low (1) Medium (2) High.

select if (educat ge 0).
select if (age ge 25 and age le 64).
select if (pinc ge 0).

cross educat by agecat.

means pinc by female by educat by agecat .

COMP PINC=PINC/100.

graph /line(multiple) =mean(pinc) By age.
```

Notes to Figure 1

- PINC fluctuates wildly with years of age. It is important to realize that these short-term fluctiation must be due to sampling variations. There is no way how income could fluctuate from one age (year) to the next.
- There is also a general trend: PINC first steeply rises with AGE and then levels off and finally starts decling with age. This is in line with theoretical expectations, in particular from ‘human capital’ theory.
- Interesting question about the trend: when does income peak?

```
regr /dep=pinc /enter=educat /SAVE=resid(res_INCOME).
```

- X1 is the residual income after the effect of EDUCAT has been controlled. This step is a bit suboptimal, but it makes it much more convenient to graph the following model implications.
- A better procedure would be to estimate the effects of EDUCAT and AGE simultaneous in the same model en plot the expected values from the part of the regression coefficients that deal with the age effects – this is much work.
- The next step is to model the trend with polynomial terms. It is common practice to center polynomial terms (=subtract the approximate age). This makes the polynomial term ‘orthogonal’, which in practice means that we can read whether an additional term has a significan effect from the t-value of the respective coefficient.
- Table 2 shows the fit statistics for the linear, quadratic and cucib effect of the age trend. The F-change show a significant increase beween a linear and quadratic model, but no further increase when we introduce a cubic term.

- Figure 2 shows how expected values from the polynomial model fit the income values (controlling education). Note that the linear model is still an approximation of the quadratic model, despite the different expectation. The quadratic trend peaks at around age 54.

Dummy variable regression

- We can also look at non-linearities by grouping the data in categories and test whether explained variance goes up, when we use an increasing number of categories. This type of model follows the data more closely, but it allows for discontinuities (discrete jumps), which can give a highly implausible model.
- You can create your own set of dummy variables, but it also work to create variables with catagories using e.g.:

```
comp age_03cat=13.3*trunc((age-25)/13.3)+25.
comp age_05cat= 8*trunc((age-25)/08)+25.
comp age_10cat= 4*trunc((age-25)/4)+25.
comp age_20cat= 2*trunc((age-25)/2)+25.
comp age_40cat= 1*trunc((age-25)/1)+25.
```

And use these in UNIANOVA with a **by** specification.

- Notice that 40 dummies are the same as the full AGE information.

- In this case you will have to calculate your own F-change values, which I did in Table 3 using an Excel sheet (it took a while before I obtained the correct numbers). 10 categories is the optimal solution, going to 20 or 40 categories increases the explained variance, but not significantly. Notice that we see a rare occasion that adj. R2 still increases, but the increase is not statistically significant according to F-change. Almost always the information is the same.

Spline regresssion

- Spline regressions offer an interesting solution between dummy variable regression and polynomials. We do not fix the functional form (unlike with polynomials), but do not allow for discontinuities.
- Table 2a gives the results for 3 5 and 7 knots (which implies 4 6 and 8 partial regression lines). In this cases the comparions of F-change favors the regression line with 3 knots (4 trajectories). Note that we do not improve this, when using 5 knots.
- I would prefer the 4-knots model in this case, but notice that the quadratic model is essentially the same and has even a slightly higher adj. R2.

The difference between F-model and F-change

As a general test on explained variance SPSS offers an over-all F-test. This statistic tests whether *any* of the predictor variables contributes significantly to the explained variance. The H0 here is that NO predictor explains any variance in the dependent variable. This F-test is often not of substantive interest, because we are truly not interested in this H0.

F-tests are generally a comparison between additional explained variance relative to what would have been explained under the H0, which is the Mean Square residual. F-tests need to be compared to a F-distribution, which has 2 degrees of freedom. DF1 is the degrees of freedom of the numerator and is the number (k) of estimated effects. DF2 is the residual degrees of freedom (N-k-1). DF2 is typically large en DF1 small. If DF1 = 1 (single added effect), we have a one-degree-of-freedom test. The critical values of F(1,many) is 3.84, an important number to memorize (=1.96**2)..

In blockwise regression, SPSS regresssion offers an F-change, which measures whether the additional explanied variance by a model relative to the previous one is statistically significant. F-change is very different from the difference in overall F between two models!!

Occasionally, we will have to calculate an F-change ourselves (e.g. when we use UNIANOVA, which has no blockwise feature). This is easy enough: F-change= (SSa-SSb)/(DFa-DFb) /MS-resid. With a-b and N-a-1 degrees of freedom. DJT gives a formula that uses R2 as ingredients.

SELECT IF and DO IF

I observed that there is some confusion of the use of SELECT IF and DO IF in SPSS. These commands share the same syntax of logical conditions and look alike:

```
SELECT IF (logical conditions).
.. data that do not meet conditions are deleter

DO IF (logical conditions).
.. data are transformed (COMPUTE / RECODE) if they meet the conditions
END IF.
```

If there is only one tranformation, the DO IF can be abbreviated to a single IF statement:

```
IF (age ge 20) age20=age-20.
```

Is the same as:

```
DO IF (logical conditions).
COMP age20=age-20.
END IF.
```

However, selecting cases and conditional data-transformations are quite different things!

Table 2: Test of non-linear trends with polynomial terms

Trend	Adj. R2	SS-model	F Change	df1	df2	p
linear	2.47%	5185.416	49.925	1	1930	.000
quadratic	3.34%	7078.052	18.386	1	1929	.000
cubic	3.34%	7169.255	.886	1	1928	.347

Total SS: 205643.

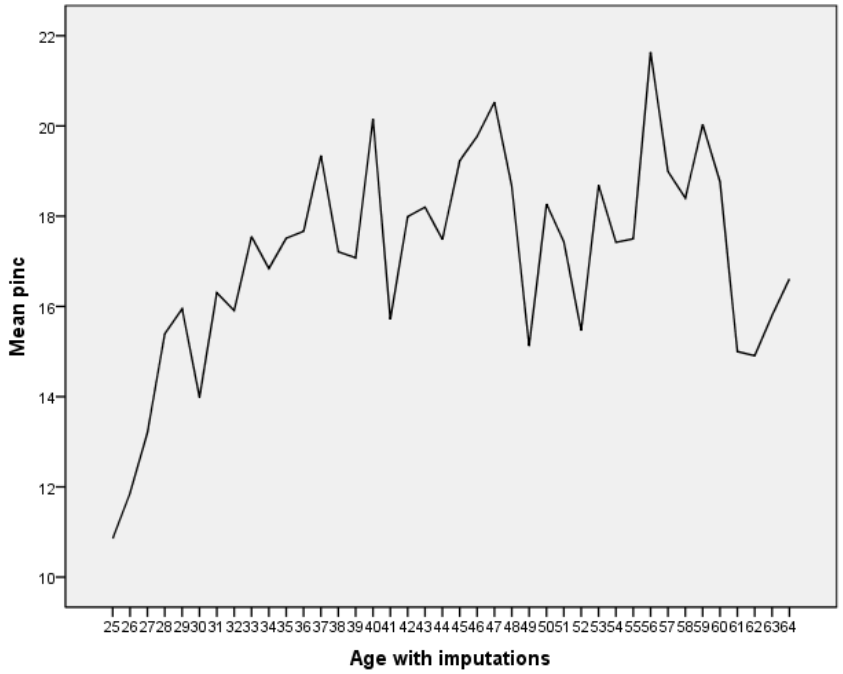
Table 3a: Test of Non-Linearity using Dummy Regression

## cat	SS-model	DF	R2	Adj R2	F-change	P
03 cat	4904	2	2.38%	2.28%		
05 cat	6290	4	3.06%	2.86%	6.75	.000
10 cat	8440	9	4.10%	3.66%	4.19	.000
20 cat	9516	19	4.63%	3.68%	1.05	>.05
40 cat	11470	39	5.58%	3.63%	0.95	>.05
SS-total= 205643						

Table 3b: Test of Non-Linearity using Dummy Regression

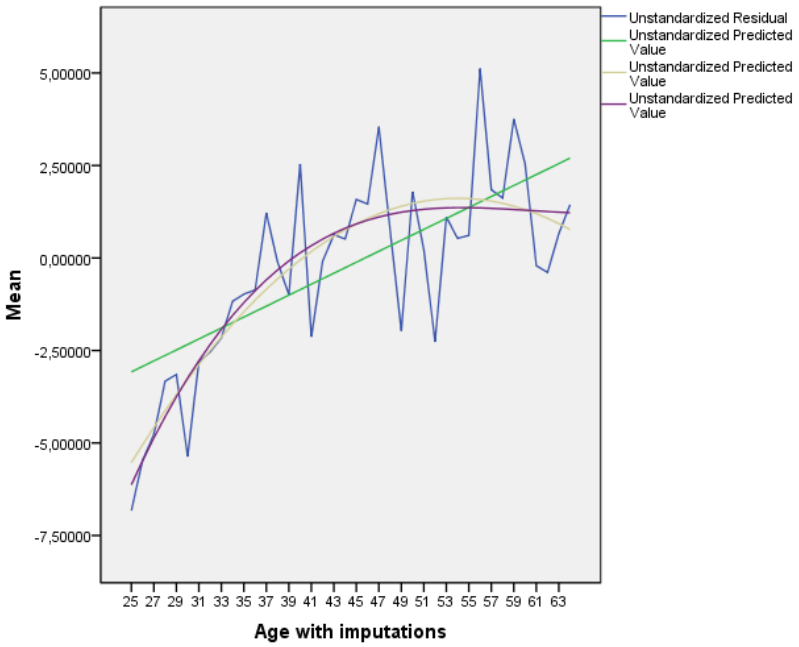
## cat	SS-model	DF	R2	Adj R2	F-change	P
03 cat	4904	2	2.38%	2.28%		
4 knots	7214	4	3.51%	3.31%	11.26	.000
6 knots	7301	6	3.55%	3.25%	0.42	>.05
8 knots	7754	8	3.77%	3.37%	2.21	>.05
SS-total= 205643						

Figure 1: Mean Income (euri per month / 100) at ages 25-64



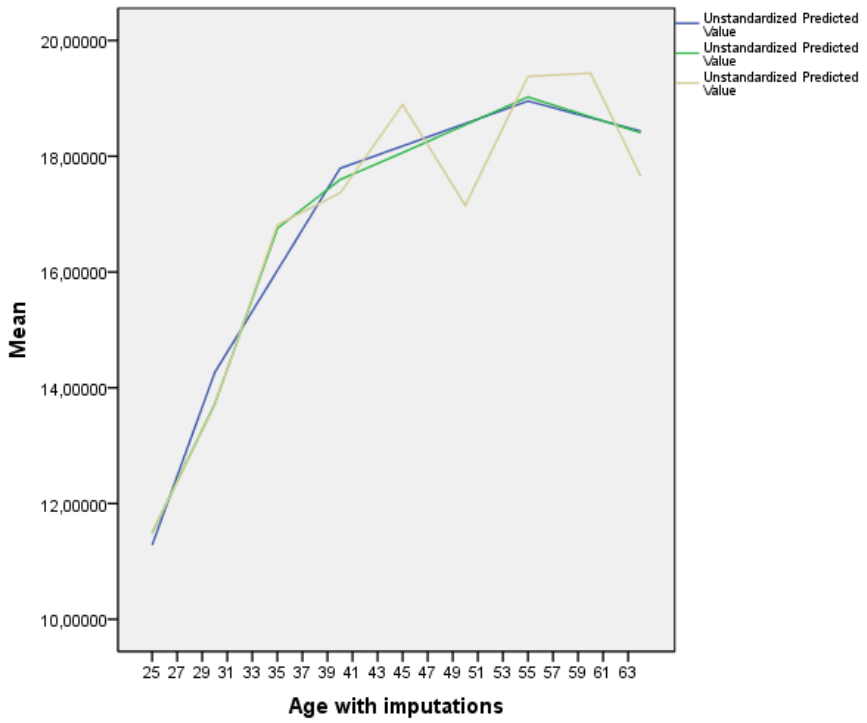
Source: ISSP-NL 2007/2008, N=1932

Figure 2: Polynomial trends in income (euri per month / 100) at ages 25-64



Source: ISSP-NL 2007/2008, N=1932. Income values are adjusted for education (3 categories).

Figure 3: Linear spline trends in income (euri per month / 100) at ages 25-64



Source: ISSP-NL 2007/2008, N=1932. Income values are adjusted for education (3 categories).

SPSS SYNTAX

```
GET
  FILE='U:\)Research\ISSP20072008\issp_2007_2008nl_def.sav'.

desc age gender z06a1 z34a.

freq z06a1 z34a.

recode z34a
(1=0)(2=400)(3=800)(4=1150)(5=1450)(6=1750)(7=2150)(8=2700)(9=3500)(10=4500)
(11=5500)
  into pinc.
recode age (25 thru 34=30)(35 thru 44=40)(45 thru 54=50)(55 thru 64=60) into
agecat.
recode z06a1 (1 2 3 6=0)(4 5 7=1)(8 9=2) into educat.
add value labels educat (0) Low (1) Medium (2) High.

select if (educat ge 0).
select if (age ge 25 and age le 64).
select if (pinc ge 0).

cross educat by agecat.

means pinc by female by educat by agecat .

** NONLINEAR REGRESSION **.

COMP PINC=PINC/100.

regr /dep=pinc /enter=educat /SAVE=resid(res_income).

mean res_income by educat agecat.

desc pinc.

desc educat /save.

freq Zeducat.

regr /dep=pinc /enter=Zeducat /SAVE=resid(res2_income).

mean res2_income by educat agecat.

comp res2_income=res2_income + 17.37.

mean res2_income by agecat.

** CREATING POLYNOMIAL TERMS **.

comp age2=(age-40)**2.
COMP age3=(AGE-40)**3.


graph /line(multiple) =mean(pinc) MEAN(X1) By age.

regr /dep=res2_income /enter=age /save=pred(linear).
regr /dep=res2_income /enter=age age2 /save=pred(QUAD).
regr /dep=res2_income /enter=age age2 AGE3 /save=pred(CUBIC).

graph /line(multiple) =mean(res2_income) MEAN(linear) MEAN(QUAD) MEAN(CUBIC)
By age.

regr /stat=def change /dep=res2_income /enter=age /enter=age2 /enter=age3 .

means pinc quad cubic by age /cells=mean.
```

```
** DUMMY VARIABLES REGRESSION **.

comp age_03cat=13*trunc((age-25)/13.3)+25.
comp age_05cat= 8*trunc((age-25)/08)+25.
comp age_10cat= 4*trunc((age-25)/4)+25.
comp age_20cat= 2*trunc((age-25)/2)+25.
comp age_40cat= 1*trunc((age-25)/1)+25.

freq age_03cat age_05cat age_10cat age_20cat age_40cat.

unianova res2_income by age_03cat
  /design=age_03cat
  /print=parameter /save=pred(x_income_03cat).

unianova res2_income by age_05cat
  /design=age_05cat
  /print=parameter /save=pred(x_income_05cat).

unianova res2_income by age_10cat
  /design=age_10cat
  /print=parameter /save=pred(x_income_10cat).

unianova res2_income by age_20cat
  /design=age_20cat
  /print=parameter /save=pred(x_income_20cat).

unianova res2_income by age
  /design=age
  /print=parameter .

graph /line(multiple) =mean(res2_income) MEAN(x_income_03cat)
  MEAN(x_income_05cat) MEAN(x_income_10cat) MEAN(x_income_20cat)
  by age .

graph /line(multiple) =mean(res2_income)  MEAN(x_income_10cat)
MEAN(x_income_20cat)
  by age .


** SPLINE REGRESSION **.

comp age_4_1=age.
comp age_4_2=0.
comp age_4_3=0.
comp age_4_4=0.
if (age ge 30) age_4_2=age-30.
if (age ge 40) age_4_3=age-40.
if (age ge 55) age_4_4=age-55.

graph /line(multiple) = mean (age_4_3) mean(age_4_2) mean(age_4_1) by age.

regr /dep=res2_income / enter=age_4_1 to age_4_4 /save=pred(ppp_4).

** SPLINE REGRESSION **.

comp age_6_1=age.
comp age_6_2=0.
comp age_6_3=0.
comp age_6_4=0.
comp age_6_5=0.
comp age_6_6=0.
if (age ge 30) age_6_2=age-30.
if (age ge 35) age_6_3=age-35.
if (age ge 40) age_6_4=age-40.
if (age ge 45) age_6_5=age-45.
if (age ge 55) age_6_6=age-55.

graph /line(multiple) = mean (age_4_3) mean(age_4_2) mean(age_4_1) by age.
```

```
delete var ppp_6.

regr /dep=res2_income / enter=age_6_1 to age_6_6 /save=pred(ppp_6).

comp age_8_1=age.
comp age_8_2=0.
comp age_8_3=0.
comp age_8_4=0.
comp age_8_5=0.
comp age_8_6=0.
comp age_8_7=0.
comp age_8_8=0.

if (age ge 30) age_8_2=age-30.
if (age ge 35) age_8_3=age-35.
if (age ge 40) age_8_4=age-40.
if (age ge 45) age_8_5=age-45.
if (age ge 50) age_8_6=age-50.
if (age ge 55) age_8_7=age-55.
if (age ge 60) age_8_8=age-60.

regr /dep=res2_income / enter=age_8_1 to age_8_8 /save=pred(ppp_8).

graph /line(MULTIPLE)=mean(ppp_4) mean(ppp_6) mean(ppp_8)  by age.
```