

## Number of children among women in Tanzania: Regression exercise no. 1

The purpose of this exercise is to illustrate the use of regression analysis when a data set with individual data is analysed. We will use data for 1000 women in Tanzania, taken from the Demographic and Health Survey (DHS) that was carried out in the 1990s in Tanzania. You will find the data at the course's website. Open the file, and store it (with a different name, if you prefer to do so) in your own directory. Next, close your Internet browser, start up Excel, and open the file.

The file contains only a few variables, selected from the DHS data set for the purpose of this exercise. In columns A-F you see the number of children the woman has given birth to, age of the woman, whether or not she was Muslim (yes=1, no=0), whether or not contraception was available within a reasonable distance (yes=1, no=0), whether or not the woman had at least some primary education (yes=1, no=0), and whether she lived in an urban (1) or rural (2) area. All variables are measured at the time of the interview.

We will attempt to explain the number of children a woman has at interview date.

In Excel 2010 you do a regression analysis by clicking "Data", "Data Analysis" and next "Regression". In case "Data Analysis" does not show up, this function has not yet been installed. To install it, you click "File", "Options" and "Add-ins". In the window that appears you should find "Excel Add-ins" at the bottom. Click "Go" and tick the box for "Analysis Toolpak".

Click "Data", "Data Analysis", and "Regression". Y is the dependent variable, X is the independent variable. In case the first row of the Y and X columns includes labels, you tick the "labels" box. Tick also "New worksheet ply". In that case, each time you do a new regression analysis, the results will be given in a new worksheet.

1. Select the number of children as the dependent variable, and age as the independent one. Estimate the constant and the coefficient in a linear model. Are the estimates significantly different from zero (at the 5% level)? What is the unit of measurement for the constant? How do you explain that the estimate for the constant is negative? Is the sign of the estimated coefficient in conformity with what you expected? Interpret the coefficient's estimate. How long is the interval between two subsequent births, on average?

$$\begin{array}{l} \# \text{ children} = -4.0 + 0.24 * \text{age} \quad R^2 = 0.625 \quad n = 1000 \\ \text{t values:} \quad \quad \quad (-23.0) \quad (40.8) \end{array}$$

unit of measurement is # children per woman.

2. Add contraception (column C) as an extra independent variable. Is there a significant effect of contraception availability on the number of children? Why is the estimated coefficient for contraception negative?

$$\begin{array}{l} \# \text{ children} = -3.83 + 0.24 * \text{age} - 0.44 * \text{contrac} \quad R^2 = 0.630 \quad n=1000 \\ \text{t values} \quad \quad \quad (-21.5) \quad (40.9) \quad \quad (-3.9) \end{array}$$

3. Add urban/rural as an extra independent variable. Explain why the estimated effect is positive. Explain why the effect of contraception no longer is significantly different from zero at the 5 per cent level.

$$\begin{array}{l} \# \text{ children} = -4.4 + 0.23 * \text{age} - 0.25 * \text{contrac.} + 0.32 \text{ urban/rural} \quad R^2 = 0.633 \quad n = 1000 \\ t \text{ values} \quad \quad (-14.5) \quad (40.8) \quad \quad (-1.77) \quad \quad (2.3) \end{array}$$

4. Check that to be Muslim or not does not have a significant effect on the number of children, once you already have included age, urban/rural, and education as independent variables. Can you think of a possible explanation?

5. The woman's educational level is given in column U, coded from 0 to 5, as indicated. Since educational level defined this way is a nominal (sometimes called "categorical") variable, it cannot be included as such as an independent variable in the regression analysis<sup>1</sup>. (Explain why.) Instead of the 6-level ordinal variable in column U, I have defined 6-1 = 5 dummy variables in columns V-Z. Check that each dummy variable corresponds to a certain level for the nominal variable. The reference category is "No education", or 0 for the ordinal variable.

Insert six blank columns next to column B (age). Move the urban/rural variable to column C, and the five dummy variables IncPrim, ..., Higher to columns D-H.

Estimate the constant and the coefficients in a linear model, which explains the number of children as a function of age, urban/rural, and the five dummy variables for educational level. What is the effect of having higher education, compared to no education at all? And how large is the effect of having higher education, compared to completed primary education? Explain, based on the estimates for the educational dummies, why the 6-level nominal variable for education would have given us incorrect estimates for the effect of educational level on the number of children.

	<i>Coefficients</i>	<i>t Stat</i>
Intercept	-4.03	-13.7
age	0.23	39.6
urb-1/rur-2	0.26	2.3
IncPrim	0.09	0.5
ComPrim	-0.40	-3.0
IncSec	-1.51	-5.4
ComSec	-1.32	-1.8
Higher	-2.40	-2.1

$$R^2 = 0.648 \quad n = 1000$$

<sup>1</sup> A nominal/categorical variable with logically ordered categories is sometimes called an ordinal variable.