ulm university universität
uulm

Universität Ulm | 89069 Ulm | Germany

Fakultät für Mathematik und Wirtschaftswissenschaften

M.Sc. Zein Kasrin
Institut für Wirtschaftspolitik

Ludwig-Erhard-Stiftungsprofessur

Sommersemester 2017

Übung zur Empirischen Wirtschaftsforschung

IX. Einkommensfunktion

Literatur

Kennedy, P.(2003), A Guide to Econometrics, Fifth Edition, Chapter. 17

Goebel, J. (2009), Introduction to the Socio-Economic Panel (SOEP), IZA Red Cube Seminar

Polachek W. (2007), Earnings Over the Lifecycle: The Mincer Earnings Function and Its Applications. IZA discussion paper 3181, Bonn

Franz, W.(2009), Arbeitsmarktökonomik, 7. Auflage, Springer

5.1 Structure of Data: Time Series, Cross-Sectional & Panel

Time Series data (Zeitreihendaten) are data collected on a *single* observational unit over *multiple* periods of time in the same method and frequency (eg., monthly, weekly, daily..etc). Usually such models explain the behavior of the dependent variable in terms of its past values. Examples of studies using time series data include:

- Time series modeling and forecasting inflation for Nigeria from 2003 to 2012
- Forecasting stock market volatility of the DAX index options market (from the 3rd of February 2002 through the 5th of December 1995).

Why use Time series data?

- To develop forecasting models. For example, based on the monthly inflation rate value for the last 5 months, what will be the inflation rate next month? Or based on the daily DAX index value for the last week, what will be the DAX index value tomorrow?.
- To estimate causal effects of an economic decision on a variable of interest. For example, if the European Central Bank (ECB) increases the interest rates today, what will be the effect on inflation in 6 months? in 12 months?

Notation for Time Series Data

y_t: Value of dependent variable y in time period t

In equation form:

$y_t = \beta_0 + \beta_1 \cdot y_{t-1} + \beta_2 \cdot y_{t-2} + \varepsilon_t$

where t is the time index.

Cross-Sectional data (Querschnittsdaten) contains information of multiple phenomena obtained for *multiple* observation units at a *single* period of time. So there are i observations, also called *cross section units*, for a single time point. These observations or cross section units could be firms, individuals, countries...etc. Examples of studies that use cross-sectional data include:

- The driving forces of economic growth for the OECD countries in 2000.
- The influence of education, age and family background on the participation of German females in 1991.

Why use Cross-sectional data?

- To study the effect of a group of independent variables on the variable of interest at a single point in time.
- To compare two different populations at a single point in time. For example, how does the labor market participation of German males and females vary with respect to the influence of educational level, marital status and region of residence in 2006.

Notation for Cross-Sectional Data

y_i: Value of dependent variable y for cross section unit i

In equation form: $y_i = \beta_0 + \beta_1 \cdot x_{1i} + \beta_2 \cdot x_{2i} + \varepsilon_i$

where i is not a time index, but the number of cross section unit (eg. the i-th country, firm, graduate, household, wage worker...etc). Therefore, the order of observations does not matter.

Cross-sectional data have a higher variance relative to time series data due to having observations only at one point in time. Therefore, a lower value of R^2 for cross-sectional data is expected.

Panel data also called **longitudinal data**, are data collected for *multiple* observation units over *multiple* periods of time. In turn, panel data is cross-sectional data over different time periods. Because a panel data structure has the characteristics of both cross-sectional data and time series data, it is the richest type of data structure. In turn the researcher is able to assess how the influence of the independent variables changes over time, and therefore more accurately determine a cause and effect relationship relative to a a time series or a cross-sectional study. The German Socieconomic Panel (SOEP) is a very prominent example of a panel data structure and will be described in Section 5.4. Examples of studies using panel data sets include:

- The driving forces of economic growth for the OECD countries. A Panel Analysis from 2000 to 2010.
- The influence of education, age and family background on the participation of German females from 1990 to 2010.

Notation for Panel Data

y_{it}: Value of dependent variable y for observation or cross section unit i at time period t

In equation form: $y_{it} = \beta_0 + \beta_1 \cdot x_{it} + \beta_2 \cdot x_{2it} + \varepsilon_{it}$

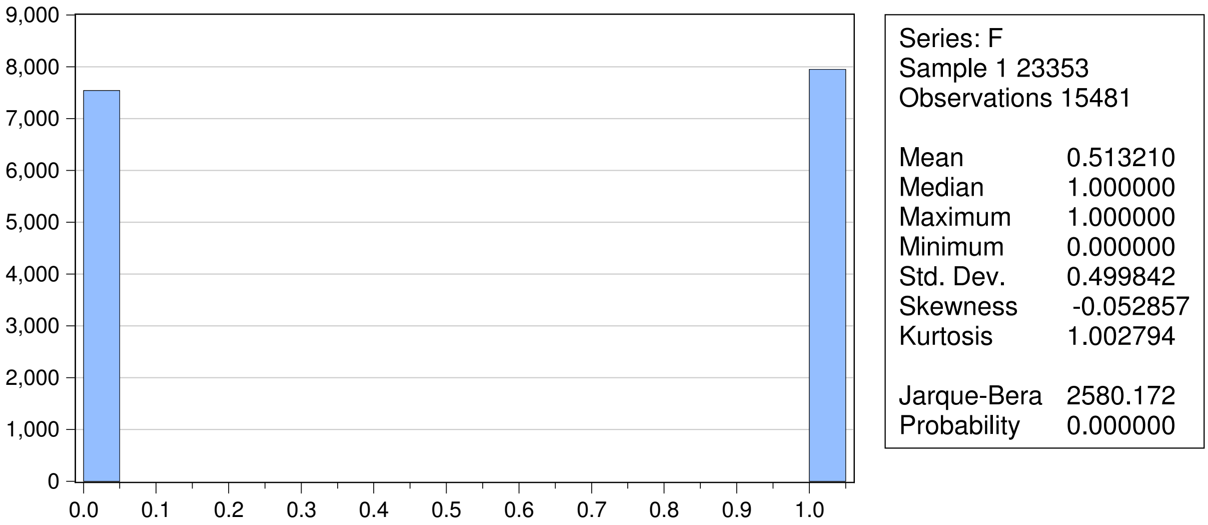
where i is the observation or cross section unit. For eg, the i-th firm, individual, country...etc, and t is the time period.

5.2 Dummy variables

Dummy variables are used to measure qualitative characteristics that are only measurable by a signal of whether a characteristic is present or absent. In turn, the dummy variable takes the value of '1' if the characteristic is present and the value of '0' if the characteristic is absent. Dummy variables have two types:

1) Dummies with two categories

- To represent Gender, eg., '1' is Female and '0' is Male.
- To represent a certain characteristic, eg., '1' is Smoker and '0' is Nonsmoker; '1' is Married and '0' is Not Married; '1' is Voted and '0' is Did not Vote; '1' is Before the War and '0' is After the War,....etc.
- The histogram below describes the dummy variable F.



```
Series: F
Sample 1 23353
Observations 15481

Mean          0.513210
Median        1.000000
Maximum       1.000000
Minimum       0.000000
Std. Dev.     0.499842
Skewness     -0.052857
Kurtosis      1.002794

Jarque-Bera   2580.172
Probability   0.000000
```

- The maximum value is 1, showing that the person is a Female
- The minimum value is 0, showing that the person is a Male
- The arithmetic mean shows that 51% of the 15,481 observations are females. Consequently, 49% of the sample are males.
- The difference between the observations present in whole sample (23,353) and those used in the histogram (15,481) reflect the presence of NAs in the gender variable.
- The median is 1, showing that there are more females than males in the sample

2)Dummies with more than two categories (Categorical dummy)

- Dummies to represent Educational Level Attainment. For example if we want to know the respondents *highest* educational attainment, we will have the following categories: *Primary Certificate, Preparatory Certificate, Secondary Certificate and University Certificate*. Each individual will have a value of '1' for only one of the categories, and the value of '0' for all other categories. We can also ask the respondents to specify *all* the educational categories that they earned. In this case, the value will be '1' for all the categories that a given individual earned, and '0' for the rest of the categories.
- Regional Dummies to represent the region an individual is living in: eg., *East Germany, West Germany, South Germany and North Germany*. Each individual will have a value of '1' for the region he is living in and a value of '0' for all other regions.
- Seasonal Dummies to represent the season of the year: *Spring, Summer, Winter and Fall*, where the value is '1' for the season in which the observation is in, and '0' for the other seasons.
- ...

To represent categorical dummy variables in a regression equation, we always include the number of categories minus one category. The category which is not included will be used as the reference category and its value will be captured in the constant of the regression equation. The value of all other dummy variables will be interpreted in comparison to the reference category. *On what basis should the researcher choose the reference category?*

Example: We have cross sectional data about the earnings of university graduates (Y_i) in 2006 based on their occupations, represented by a categorical dummy with three categories: Doctors (D_D), Lawyers(D_L) and Professors(D_P). Writing this in an equation form, we must *only* include two of the three categories, for example we include (D_D) and (D_L):

$Y_i = \beta_0 + \gamma_1 \cdot D_{D_i} + \gamma_2 \cdot D_{L_i} + \varepsilon_i,$

where;

- Y_i is the income of university graduate *i*,
- γ_1 is the difference in income between *Doctors* and *Professors*.
- D_{D_i} takes the value of '1' if individual i is a *Doctor* and the value of '0' otherwise.
- γ_2 is the difference in income between *Lawyers* and *Professors*.
- D_{L_i} takes the value of '1' if individual i is a *Lawyer* and the value of '0' otherwise.
- β_0 shows the base income. When both (D_D) and (D_L) are zero, it shows the income of *Professors*,

Please calculate the income level for each occupation:

Doctors:
Lawyers:
Professors:

Usually continuous variables are also included in the regression equation with dummies, such as the years of Experience of an individual, Age of an individual..etc.

5.3 Human Capital Theory

Human capital corresponds to any stock of knowledge or characteristics that the worker has which contributes to his productivity. The modern theory of human capital was developed by Jacob Mincer, Theodore Schultz and Gary Becker. The theory explained that the neoclassical assumption of homogeneous human input in the production function is not accurate. Indeed, individuals differ in their abilities, which means that the productivity of each individual is not identical. Consequently, individuals with higher abilities will contribute more to output. Jacob Mincer's innovation was to empirically measure human capital by treating schooling and training as investment opportunities which individuals invest in to maximize their future returns. The basic Mincer earnings function is estimated for a cross section using the following regression equation:

$$log(y_i) = \beta_0 + \beta_1 Schooling_i + \beta_2 Experience_i + \beta_3 Experience_i^2 + \varepsilon_i \tag{1}$$

Where *Schooling* refers to the years of schooling of individual *i* and *Experience* refers to the years of actual or potential experience (measured by age-years of schooling-6) of individual *i*. The coefficients from the OLS regression of figure (1) are treated as showing the individual's private returns to schooling and experience. The value of the coefficients shows us the percentage change in the logarithmic income as the coefficients change, which is equal to the percentage change in income. For example, if the estimated value for β_1 is 0.2, what does this mean economically?

- It means that having an additional year of schooling allows the individual to earn 20% $(\beta_1 * 100)$ more wage than the wage he would earn without this additional year. If we assume that the forgone wages are his only current cost, then by giving up 100% of his wages now, he can earn 20% more in all subsequent years. If our schooling variable was a set of dummies, then a similar interpretation can be made.

Social returns of schooling (as opposed to private returns) allow us to estimate how much society gains from an individual's schooling. As previously mentioned higher educated persons can contribute more to output and growth. However, this requires us to also calculate the cost of tuition and anything else paid by the government. We focus on estimating only private returns in this class.

The experience squared term in equation (1) is also estimated to capture the phenomenon that returns to experience are not stable over time. For example, having one year of experience provides a different return to having five years of experience, and so on. In order to give this flexibility in our estimations, we need to also model the squared term for experience, as the linear experience term does not change with respect to experience. Usually we will get a positive value for the linear term (β_2) and a negative value for the polynomial term (β_3), which shows that returns first increase with years of experience, but then the rate of increase decreases, and can become negative. To analyze the effect of a change of experience on output:

$$\Delta ln(y) = (\beta_2 + 2\beta_3 Experience) \tag{2}$$

We usually show the estimation results for the experience in a graphical plot to allow for visualization.

Reasons for falling returns to experience is that earnings usually rise as workers grow older and more experienced, until reaching a certain prime age where returns reach a peak. After that, additional years of experiences do not lead to higher income as people become older and less productive.

Labor economists have intensively used the Mincer equation and included other variables to explain earnings such as gender, region of residence, sector of work, ethnic differences..etc.

We will use a cross-section of the German Socio-Economic Panel (GSOEP) to estimate human capital earning functions for 2012. First we provide a short description of SOEP in the next section.

5.4 The SOEP Panel Survey

Das Sozio-Ökonomische Panel (SOEP) is a representative longitudinal survey of private households and persons in Germany which was started in 1984. The survey is performed annually, and tracks around 11,000 households (HH) and 30,000 persons. The objective of the panel study is to collect representative micro-data on the sample in order to measure changes in living conditions and quality of life, in addition to the sample's opinions on various social and political issues. The survey includes a household questionnaire and an individual questionnaire. The essential areas of interest of the study are:

- Satisfaction with life in general and in certain life aspects
- Family situation and background
- Characteristics and facilities of place of residence
- Education and training level
- Employment status and occupation
- Income sources and working hours
- Health Situation
- Personal assets and liabilities
- Attitudes and opinions in social and political aspects

SOEP contains different sample groups divided as follows:

- **A Deutsche (West)**. Started in 1984, includes 4528 HH. Head is either German or other nationality than those in Sample B.
- **B Ausländer (West)**. Started in 1984, includes 1393 HH. Head is either Turkish, Italian, Spanish, Greek or Yugoslavian (i.e. main foreigner groups of 'guestworkers').
- **C Deutsche (Ost)**. Started in 1990, includes 2179 HH. Head was a citizen of GDR.
- **D Zuwanderer 1984-93**. Started in 1994, includes 522 HH. It refers to HH where at least one HH member moved to Germany after 1984.
- **E Ergänzung 1998**. Includes 1067 HH. Random sample covering all existing subsamples.
- **F Ergänzung 2000**. Includes 6052 HH. Random sample covering all existing subsamples.
- **G Hohe Einkommen 2002**. Includes 1224 HH. It covers HH whose monthly net income $>$ 7,500 DM.

- **H Ergänzung 2006**. Includes 1506 HH. Random sample covering all existing subsamples.
- **J Aufstockung 2011**.
- **K Aufstockung 2012**.

The SOEP questionnaires and variables can be viewed on `https://data.soep.de`. In the workfile `Uebung8.wf1` is a collection of chosen variables from the SOEP panel survey of 2012. The sample size is restricted by 25% by means of a random variable due to Datenschutz.

Data in Workfile Uebung9.wf1

Y	*Bruttomonatseinkommen in €*
F	*Geschlecht, Dummyvariable, 1 für Frauen, 0 für Männer*
SCH	*Schulabschluss*
HAUPTORLESS	*Hauptschulabschluss oder weniger*
REAL	*Realschulabschluss (Referenzgruppe)*
ABI	*(Fach-)Hochschulreife*
OHNEAUSBILDUNG	*ohne Ausbildungsabschluss*
BBILDUNG	*abgeschlossener Berufsausbildung (Referenzgruppe)*
UNI	*(Fach-)Hochschulabschluss*
XYR	*Berufserfahrung (in Jahren)*

Quelle: Ausgewählte Daten des SOEP für das Jahr 2012.

5.5 Estimating the Earnings Function using SOEP 2012

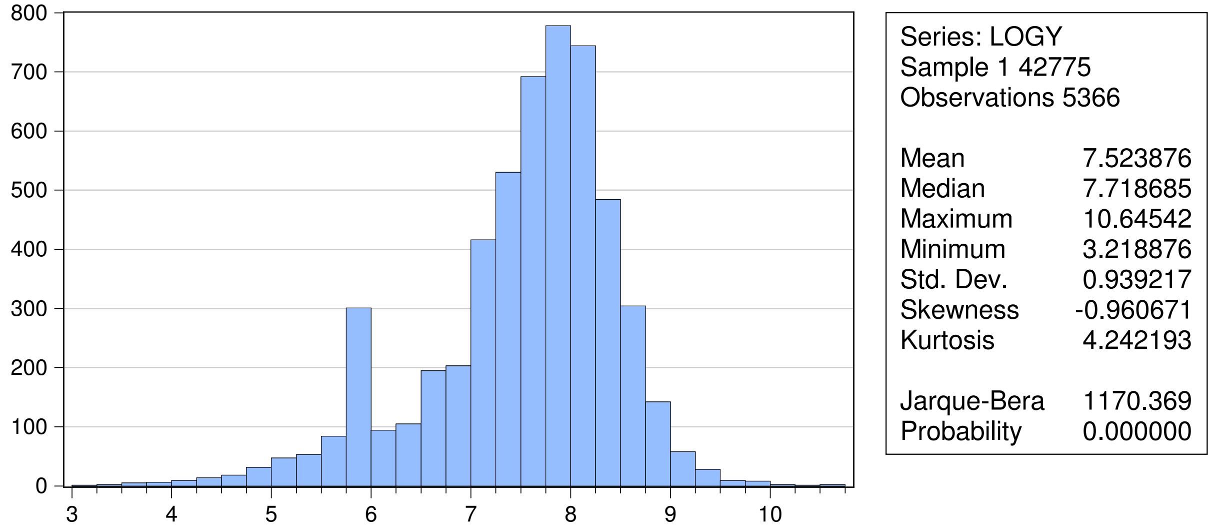
We will estimate a semilog Mincer's earnings functions, as explained in the human capital theory. The equation is derived from the multiplicative model given by:

$y = e^{\beta_0} \cdot e^{\beta_1 x_1} \cdot e^{\beta_2 x_2} \cdot e^{\beta_3 x_3} \cdot e^{\varepsilon}$

This translates to:

$$\begin{aligned} log(Y)_i = \beta_0 + \beta_1 \cdot XYR_i + \beta_2 \cdot XYR_i^2 + \beta_3 \cdot HAUPTORLESS_i + \beta_4 \cdot ABI_i \\ + \beta_5 \cdot KEINAUSBILDUNG_i + \beta_6 \cdot UNI_i + \beta_7 \cdot F_i + \varepsilon \end{aligned}$$

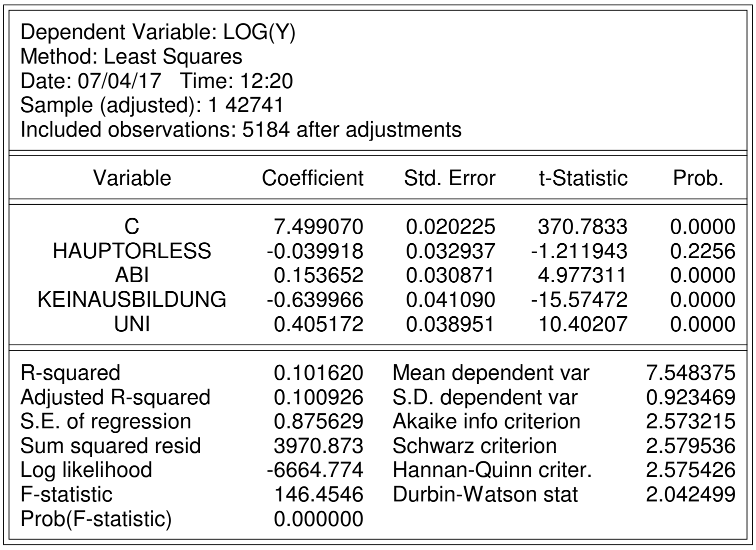
We can calculate the value of the income by taking the antilog, i.e. $e^{(\beta)}$.



```
Series: LOGY
Sample 1 42775
Observations 5366

Mean          7.523876
Median        7.718685
Maximum       10.64542
Minimum       3.218876
Std. Dev.     0.939217
Skewness     -0.960671
Kurtosis      4.242193

Jarque-Bera   1170.369
Probability   0.000000
```

- 5,366 individuals from the total sample size of 42,775 reported an income.
- The mean monthly income is $e^{7.52}$= 1,844 €.
- The distribution is left skewed, which means that the mass of the distribution is concentrated on the right of the figure.
- The maximum monthly income value is 41,772 € $(e^{10.64})$ and the minimum monthly income value is 25 € $(e^{3.21})$

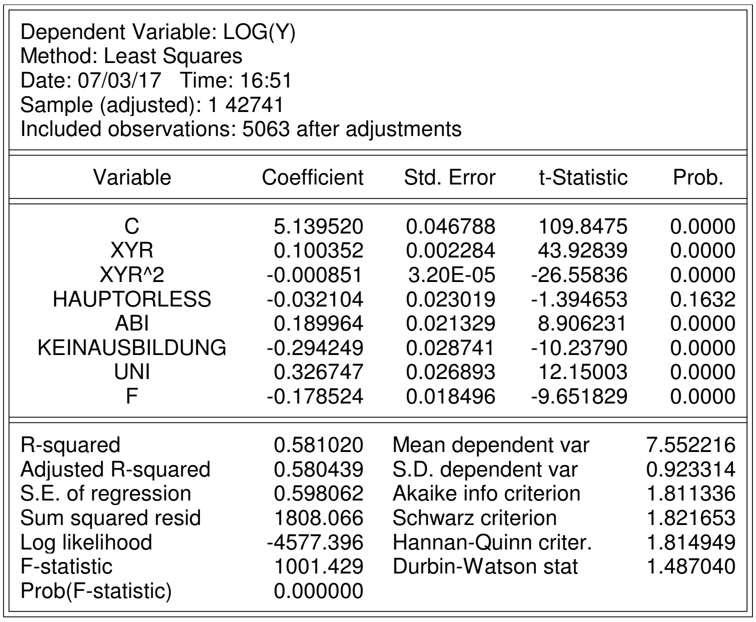
We will start by estimating an earnings function with only the educational categories dummies as the independent variables:



```
Dependent Variable: LOG(Y)
Method: Least Squares
Date: 07/04/17   Time: 12:20
Sample (adjusted): 1 42741
Included observations: 5184 after adjustments

        Variable          Coefficient   Std. Error   t-Statistic    Prob.

           C               7.499070     0.020225     370.7833     0.0000
      HAUPTORLESS         -0.039918     0.032937    -1.211943     0.2256
          ABI              0.153652     0.030871     4.977311     0.0000
    KEINAUSBILDUNG        -0.639966     0.041090    -15.57472     0.0000
          UNI              0.405172     0.038951     10.40207     0.0000

R-squared                  0.101620   Mean dependent var       7.548375
Adjusted R-squared         0.100926   S.D. dependent var       0.923469
S.E. of regression         0.875629   Akaike info criterion    2.573215
Sum squared resid          3970.873   Schwarz criterion        2.579536
Log likelihood            -6664.774   Hannan-Quinn criter.     2.575426
F-statistic                146.4546   Durbin-Watson stat       2.042499
Prob(F-statistic)          0.000000
```

- The estimation results show that all educational dummies are statistically significant at the 1% significance level, except for HAUPTORLESS, what does this mean?.
- The results for each group of dummies will be compared to their reference group. So the results for Schulabschluss dummies is compared to REAL. Similarly, the results for the Ausbildungabschluss dummies will be compared to BBILDUNG.
- The value of the constant C shows the average earnings of a person with the reference education category. So this is a person with REAL or REAL and BBILDUNG abschluss. This person earns $e^{7.49}$ which equals 1790 €.
- The UNI coefficient shows that those with Uniabschluss receive the highest earnings, as they receive 40% more earnings relative to those with Berufsausbildung.
- Similarily, persons whose highest education level is ABI earn 15% more as those with REAL.
- How much does a person with ABI and UNI earn? The person earns $e^{7.49+0.15+0.40}$ which equals 3,102 €.
- What does the value of the constant show in this estimation?.

The results from the estimation show that the income level is higher with higher education levels. However, the R^2 value of the estimation shows that only 10.1% of the variation in income is explained by our model. This is because other important human capital variables that influences earnings were not included in the model. We now estimate the full human capital equation which we showed in the beginning of this section, by adding experience and the gender dummy:



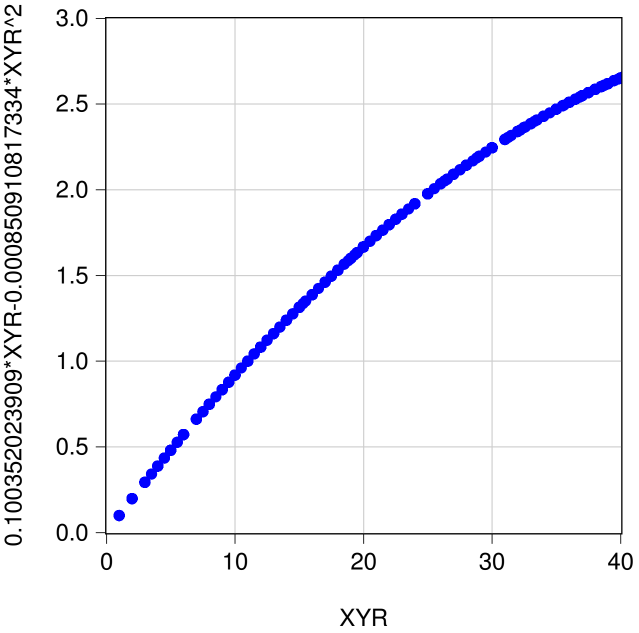
```
Dependent Variable: LOG(Y)
Method: Least Squares
Date: 07/03/17   Time: 16:51
Sample (adjusted): 1 42741
Included observations: 5063 after adjustments

        Variable        Coefficient    Std. Error    t-Statistic    Prob.

            C             5.139520     0.046788      109.8475      0.0000
           XYR            0.100352     0.002284      43.92839      0.0000
          XYR^2          -0.000851     3.20E-05     -26.55836      0.0000
      HAUPTORLESS        -0.032104     0.023019     -1.394653      0.1632
           ABI            0.189964     0.021329      8.906231      0.0000
    KEINAUSBILDUNG       -0.294249     0.028741     -10.23790      0.0000
           UNI            0.326747     0.026893      12.15003      0.0000
            F            -0.178524     0.018496     -9.651829      0.0000

R-squared               0.581020   Mean dependent var      7.552216
Adjusted R-squared      0.580439   S.D. dependent var      0.923314
S.E. of regression      0.598062   Akaike info criterion   1.811336
Sum squared resid       1808.066   Schwarz criterion       1.821653
Log likelihood         -4577.396   Hannan-Quinn criter.    1.814949
F-statistic             1001.429   Durbin-Watson stat      1.487040
Prob(F-statistic)       0.000000
```

- This regression equation has a much higher explanatory power, given by an R^2 of 58.1%. Furthermore, all coefficients are statistically significant.

- The educational dummies show a similar trend compared to the first estimation, but their influence on earnings decreased for almost all levels. *Why*?

- The dummy variable F shows that a female earns 17% less than a male with a corresponding qualification.

- The experience and experience squared coefficients are significant and have the expected signs. This shows that experience profiles of wage workers are inversely U-shaped, as explained by human capital theory. We can plot the *total effect* of experience on income to allow for graphical visualization:



The plot shows that returns to experience increase by 100% for a person having ten years of experience, but then the rate of increase falls as the years of experience increase. For example, with twenty years of experience, returns to experience increase by around 68%. With thirty years of experience, the increase is 56%, and with forty years of experience returns increase by 39%.

How to show the *marginal effect* of experience on earnings in a plot? That is, the rate of change in returns per year of experience (i.e. what we showed in equation 2). Please perform the plot in Eviews and interpret the results.