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 2015

Übung zur Empirischen Wirtschaftsforschung

IX. Einkommensfunktion

Literatur

Smolny, W., and Kirbach, M. (2004): Wage differentials between East and West Germany - Is it related to the location or to the people?

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

Helmholtzstr. 20, Raum E 02
Tel. 0731 50-24265, Fax -24262

http://www.uni-ulm.de/wipo
zein.kasrin@uni-ulm.de

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 and a few other variables. Examples include:

- Unemployment rate per quarter for Germany from 1960-2011
- Inflation rate per month for France from 1975-2011
- Value of the DAX index per day from 1993-2011
- Export rate of German products per month from 2010-2015

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 the 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 measured variable y in time period t

In equation form:

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

where t is the time index.

Cross-Sectional data (Querschnittsdaten) are data collected 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 include:

- The income and expenditures of Economics Master graduates of Ulm university in July 2012.
- Unemployment rates for Middle East and North Africa (MENA) countries in the year of 2013.

- The inflation rates for all EU countries in 2008.
- The profits and costs of German Automobile firms in Januray 2014.

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. For example, what is the effect of educational level, marital status, and region of residence on female labor market participation in Germany in 2006.
- 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 measured 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).

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 a panel data include:

- The yearly income and expenditure of Economics Master graduates of Ulm university from 2010 to 2015
- Yearly unemployment rates for Middle East and North Africa (MENA) countries from 2009 to 2014
- Quarterly inflation rates for all EU countries from the first quarter of 2008 to the last quarter of 2014
- Monthly profits of German Automobile firms from January 2010 to August 2014

Notation for Panel Data

y_{it}: Value of measured variable y for observation or cross-sectional unit i at time period t

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

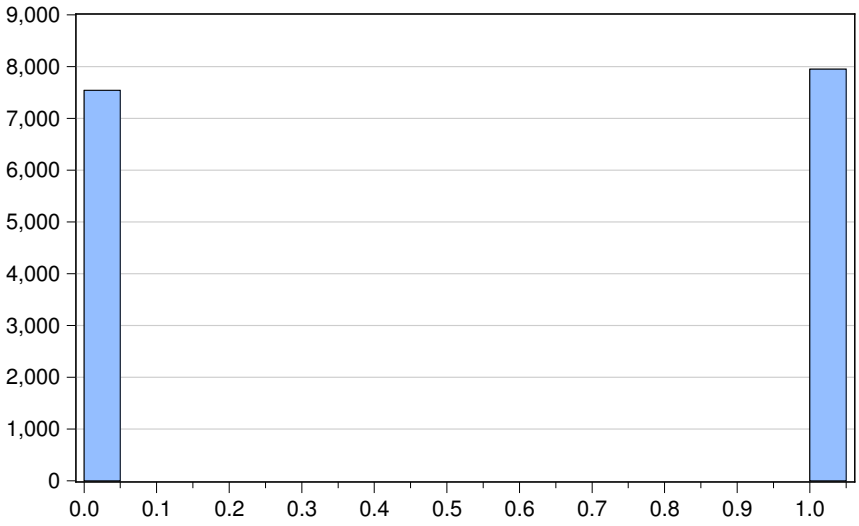
where i is the observation or cross section unit. For eg, the i-th firm, individual, country...etc, and t is the time period.

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.



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. The value of all other dummy variables will be interpreted in comparison to 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*,

In turn, we calculate the income level by occupation type as follows:

Doctors: $Y = \beta_0 + \gamma_1$
Lawyers: $Y = \beta_0 + \gamma_2$

Professors: $Y = \beta_0$

Usually other 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 man behind the modern earnings function theory is Jacob Mincer. In 1958, he presented his theory about measuring earnings which was an innovation in the labor economic field. Mincer and other authors 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, the influence of individuals on total output and income will also not be constant. Mincer's innovation was to treat schooling and training as investment opportunities which individuals invest in to maximize their future returns. The basic model was written as:

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

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 since used the Mincer equation and included other variables to explain earnings such as gender, region of residence, sector of work, ethic differences..etc.

We will estimate an earnings function for individuals using the German Socio-Economic Panel (SOEP), where we consider how investment in different human capital variables influence the income of individuals. First we provide a 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 in the website `https://data.soep.de`. In the workfile `Uebung9.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.

Data in Workfile Uebung9.wf1

Y	*Bruttomonatseinkommen in €*
F	*Geschlecht, Dummyvariable, 1 für Frauen, 0 für Männer*
OHNE	*kein Schulabschluss (Referenzgruppe)*
HAUPT	*Hauptschulabschluss*
REAL	*Realschulabschluss*
FACHSCHU	*Fachhochschulreife*
ABI	*Abitur*
KEINBERUF	*ohne Berufsabschluss (Referenzgruppe)*
LEHRE	*Lehre*
MEISTER	*Meister*
UNI	*Universitätsabschluss*
XYR	*Berufserfahrung (in Years)*
STUND	*Tatsächliche Arbeitszeit pro Woche (in Hours)*
PSAMPLE	*Stichprobenart (1: Deutsche (West), 2: Ausländer (West), 3: Deutsche (Ost), 4: Zuwanderer 1984-93, 5: Ergänzung 1998, 6: Ergänzung 2000, 7: Hocheinkommensbezieher 2002: 8: H Ergänzung 2006, 10: J Aufstockung 2011, 11: K Aufstockung 2012)*

Quelle: Ausgewählte Daten des SOEP für das Jahr 2012.

5.5 Estimating the Earnings Function using SOEP 2012

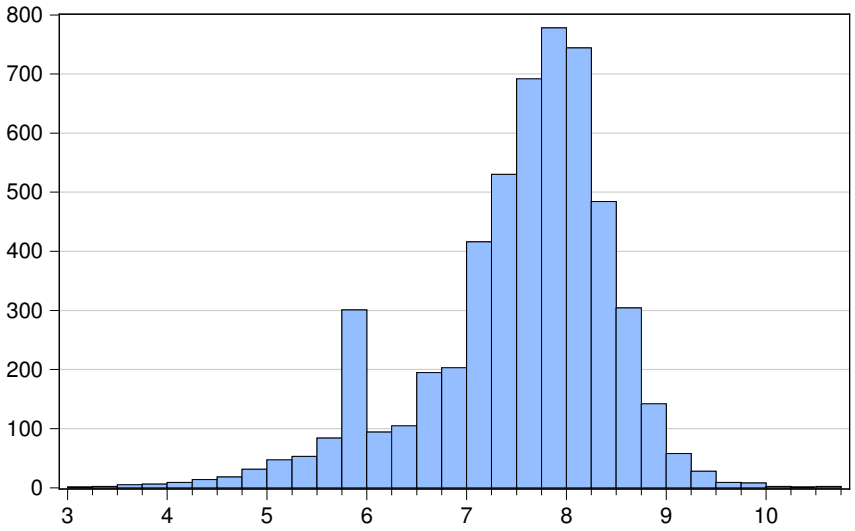
We will estimate a semilog earnings functions, as explained in the human capital theory. The equation is derived from the multiplicative model given by:

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

This translates to:

$$log(y)_i = \beta_0 + \beta_1 \cdot log(Stund)_i + \beta_2 \cdot Xyr_i + \beta_3 \cdot Xyr_i^2 + \beta_4 \cdot Haupt_i + \beta_5 \cdot Real_i + \beta_6 \cdot Fachschu_i + \beta_7 \cdot Abi_i + \beta_8 \cdot Lehre_i + \beta_9 \cdot Meister_i + \beta_{10} \cdot Uni_i + \beta_{11} \cdot F_i + \varepsilon$$

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



Series: LOG(Y)	
Sample 1 23353	
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 23,353 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. We restrict our estimations to subsamples A to C (i.e. psample<=3):

Dependent Variable: LOG(Y)
Method: Least Squares
Date: 06/24/15 Time: 17:23
Sample: 1 23353 IF PSAMPLE<=3
Included observations: 1578

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	6.787954	0.184257	36.83956	0.0000
HAUPT	0.561388	0.189449	2.963266	0.0031
REAL	0.435292	0.187903	2.316579	0.0207
FACHSCHU	0.753750	0.201946	3.732427	0.0002
ABI	0.644436	0.192308	3.351066	0.0008
LEHRE	0.186668	0.051743	3.607598	0.0003
MEISTER	0.483444	0.084584	5.715546	0.0000
UNI	0.541038	0.062257	8.690471	0.0000

R-squared	0.114273	Mean dependent var	7.611071
Adjusted R-squared	0.110324	S.D. dependent var	0.846932
S.E. of regression	0.798848	Akaike info criterion	2.393766
Sum squared resid	1001.909	Schwarz criterion	2.420959
Log likelihood	-1880.681	Hannan-Quinn criter.	2.403870
F-statistic	28.93657	Durbin-Watson stat	1.897383
Prob(F-statistic)	0.000000		

- The estimation results show that all educational dummies are statistically significant at the 1% level.
- The results for each group of dummies will be compared to their reference group. So the results for Schulabschluss dummies is compared to OHNE, the people without a Schulabschluss. Similarly, the results for the Berufabschluss dummies will be compared to KEINBERUF, and the results of Uni are compared to those with KEINUNI.
- The value of the constant coefficient, C, shows the income received for persons who are ohne Schulabschluss, und mit kein Beruf und kein Uni Abschluss. These persons receive an income of $e^{6.78}$= 880 €.
- The UNI coefficient shows that those with a UNI Abschluss receive the highest earnings, as they receive 54% more earnings relative to those without a UNI Abschluss. The next highest earnings level is MEISTER with 48% higher earnings relative to those with KEINBERUF Abschluss.
- Persons whose highest education level is Hauptabschluss earn 56% more as those OHNE Schulabschluss. In specific they earn $e^{6.78+0.56}$ or ($e^{6.78} \cdot e^{0.56}$) which equals 1,540 €. With the same logic, persons whose highest education level is Abitur earn 64.5% higher earnings than persons OHNE Schulabschluss. In specific, they earn $e^{6.78+0.64}$= 1,669 €.
- Persons with LEHRE Abschluss earn 18.7% more as those with KEINBERUF Abschluss. In specific, They earn $e^{6.78+0.18}$= 1,053 €.
- Persons with ABI and UNI earn $e^{6.78+0.64+0.54}$= 2,854 €.

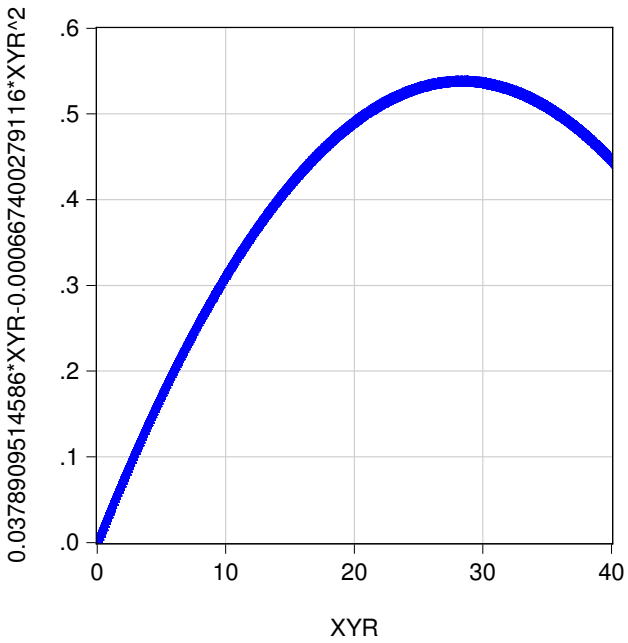
The results from the estimation show that the income level is higher with higher education levels. However, the R^2 value of the estimation is only 11.4%, which means that only 11.4% of the variation in income is explained by our model. This is because other important human capital variables that influence earnings such as experience and work hours are not included in the model. We now estimate the full human capital equation which we showed in the beginning of this section:

Dependent Variable: LOG(Y)
Method: Least Squares
Date: 06/24/15 Time: 19:47
Sample: 1 23353 IF PSAMPLE<=3
Included observations: 1466

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.340598	0.170248	19.62191	0.0000
LOG(STUND)	1.031709	0.030551	33.76994	0.0000
HAUPT	0.135554	0.133647	1.014264	0.3106
REAL	0.129343	0.133009	0.972433	0.3310
FACHSCHU	0.460823	0.140965	3.269066	0.0011
ABI	0.481487	0.135770	3.546354	0.0004
LEHRE	0.021845	0.035254	0.619660	0.5356
MEISTER	0.173792	0.056487	3.076676	0.0021
UNI	0.193532	0.042280	4.577404	0.0000
F	-0.122818	0.030012	-4.092294	0.0000
XYR	0.037891	0.004047	9.362920	0.0000
XYR^2	-0.000667	9.40E-05	-7.097811	0.0000

R-squared	0.613850	Mean dependent var	7.679791
Adjusted R-squared	0.610928	S.D. dependent var	0.804649
S.E. of regression	0.501905	Akaike info criterion	1.467339
Sum squared resid	366.2747	Schwarz criterion	1.510643
Log likelihood	-1063.559	Hannan-Quinn criter.	1.483490
F-statistic	210.1249	Durbin-Watson stat	1.641934
Prob(F-statistic)	0.000000		

- As a first glimpse this regression equation has a much higher explanatory power, given by an R^2 of 61.3%. Furthermore, all coefficients are statistically significant, except for HAUP, REAL and LEHRE. These result mean that persons whose highest education level is HAUP or REAL do not receive significantly higher earnings relative to those OHNE Schulabschluss. Similarly, those whose highest education level is LEHRE, do not receive significantly higher earnings relative to those with KEINBERUF Abschluss.
- The other educational levels show a similar trend compared to the first estimation, but their influence on earnings decreased for all levels. *Why*?
- The results for $log(STUND)$ show that when weekly work hours are 1% higher, this leads to 1% higher income.
- The results for F show that females earn 12.2% lower earnings relative to males.
- 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 are 30% 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 20%. With thirty years, the increase is only 4%, and with forty years returns decline by 9%.