Working paper
1/2005

Lifetime earnings

Fedor Iskhakov

Stiftelsen Frischsenteret for samfunnsøkonomisk forskning
Ragnar Frisch Centre for Economic Research

Working paper 1/2005

Lifetime earnings

Fedor Iskhakov

Abstract: This essay examines the problem of choice of a simple model for predicting unobserved and future earnings from pension point histories on the bases of lifecycle approach. Effects of specific data censoring technique used by the Norwegian statistics agency are carefully investigated.

Keywords: earnings prediction, pension points, censored variables models

Contact: f.v.iskhakov@frisch.uio.no, www.frisch.uio.no, phone + 47 22 95 88 37

Report from the project "Working life and welfare of the elderly" (internal project no. 1133), funded by the Norwegian Research Council

ISBN 82-7988-058-5
ISSN 1501-9241

1. Introduction

Predicting potential earnings for a sample of individuals is an essential task in many labour market studies analyzing personal behavior. When behavior is thought of as a sequence of discrete choices performed by an agent, or in other word, when at each time period individuals are choosing one out of a full set of mutually exclusive alternatives, observations usually present characteristics of no other than the chosen alternative. In order to estimate a conditional logit model in this case one has to apply some prediction method to obtain characteristics of feasible but not chosen alternatives. Earnings is one of the most important among these characteristics. Moreover, many times other important variables are defined or can be assessed through earnings, for instance occupational and social security pensions and some other social benefits. Thus, the problem of predicting potential earnings is quite important in the labour market research and requires close attention.

One of the simplest approaches to earnings prediction lies within the framework of life cycle modeling. The main theme in this type of research is the notion of individuals who develop and realize carefully planned lifetime program for most economic variables they deal with. This assumption, although quite questionable, gives theoretical background for search of repeated patterns and cycles within the life period of individuals under consideration, particularly for the similar patterns in their earnings profiles. With support of such reasoning in current paper a simple quadratic lifecycle model is developed for lifetime earnings.

However there is a considerable complication for such model which originates in the Norwegian institutional settings. On one hand, statistic authorities in Norway have beautiful record of earning histories for nearly whole population from the year 1967 when the earnings based public pension was introduced. On the other hand, this data is collected for public pension calculation, and that is done with the help of so called “pension points”. The problem is that expressed in pension points histories of earnings are censored from above and from below according to the evolving legislation due to the fact that extremely low and extremely high incomes are irrelevant for calculating public pension benefits. Thus, strictly speaking, regular regression estimation methods are not applicable in the case and censored models need to be developed and used.

In the technical note [Iskhakov, Kalvarskaia, 2003] the problem of earnings prediction was solved roughly without paying attention to the nature of the data involved. This work is used as a reference case and the results obtained there will be improved in the current paper.

The paper is organized as follows. First, data structure, its origins and collection principles are reviewed and some descriptive statistics are demonstrated. Second, several statistical models are spelled out and estimated. Finally, alternative setups are tested by the accuracy of out–of–sample predictions and compared to the straight forward classical regression approach, and the best model is nominated.

2. Description of the data

The rules for calculating pension points from annual earnings have changed twice since introduction in 1967 (details can be found in [Haugen, 2000] and [Røgeberg, 2000]). Denote I annual real pension generating income (measured in terms of base pension amount G[1]) and P corresponding pension point. Then for the period from 1967 to 1970 formula (1) was used.

$$P = \begin{cases} 0, I \le 1; \\ I - 1, 1 < I \le 8; \\ 7, I > 8. \end{cases} \tag{1}$$

In 1970 the upper censoring was altered introducing formula (2).

$$P = \begin{cases} 0, I \le 1; \\ I - 1, 1 < I \le 8; \\ \frac{13}{3} + \frac{I}{3}, 8 < I \le 12; \\ 8\frac{1}{3}, I > 12. \end{cases} \tag{2}$$

And finally in 1992 the upper censoring limit was brought back down with additional change for earnings above 6G.

$$P = \begin{cases} 0, I \le 1; \\ I - 1, 1 < I \le 6; \\ 3 + \frac{I}{3}, 6 < I \le 12; \\ 7, I > 12. \end{cases} \tag{3}$$

Chart 1 presents the function mapping annual earnings into pension point visually. Here the solid line (oABm) corresponds to the period 1967 to 1970 (regime 1), short–dashed line

[1] Basic pension G is used throughtout the paper as the main quantity measure for eanings. Since G is corrected from time to time according to macroeconomic situation we are dealing with earnings in real terms. Details and tables of G can be found in [Haugen, 2000].

(oABCn) to the period 1970 to 1992 (regime 2) and long–dashed line (oADm) to the period 1992 and onwards (regime 3).

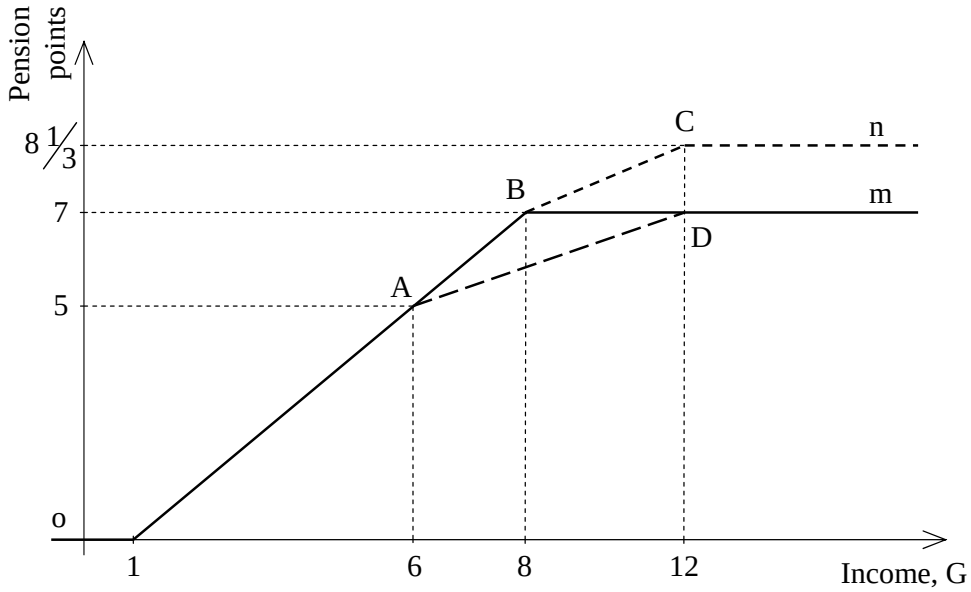

Chart 1. Mapping functions in different time intervals.

It's clearly seen from Chart 1 that the income variable is truly censored making the information on annual earnings limited. Indeed, even when assumed non–negative, incomes below 1G are mapped into P=0 pension point and in all three regimes incomes above 12G are represented by the pension points P=7 or P=8,33. Our intension is to model true incomes I with the censored data available on P.

In order to develop more or less general model for lifetime earning profiles for Norwegian individuals consider the following reasoning on sample definition. Start from pension point files that contain records with accumulated pension rights from 1967 to 2000[2] (34 points). Assume full history of earnings to be 49 points according to working ages from 19 to 67. To utilize the data the best way possible consider only those people whose earnings history fully overlaps with the observation window, this requires individuals to be at least 53 years of age in 2000 and no more than 33 in 1967. Thus, we obtain a sample of all individuals born within the time span 1934 to 1947. Additionally, consider only those individuals alive in 1993 (this

[2] The files mentioned in this paper and used to generate data for model estimation are register data gathered and provided by Statistics Norway for the use within Frisch Centre for Economic Research.

limitation follows from the demographic data availability). This sample should serve well for our purpose of finding general shape of the earning profile in spite of possible selection bias.

The defined sample contains 658 677 individuals. Among these some people still have quite short histories of earnings above 1G, apparently due to lack of jobs during most of their lives, and probably living on social benefits. Such observations were dropped with a cut off point at 20 years with positive pension points leading to about 22% reduction of the sample. We are left then with 511 911 individuals. Among these 32 224 persons turn 60 in the year 2001 and one thousand of them selected randomly will represent a target population of some specific labour market research project for which we test different models in part 3 of this paper. These individuals are set aside leaving 510 911 individuals for model estimation.

The chosen individuals altogether present 17 370 857 non–missing observations of pension points. Introduce a time index variable t taking values from 1 to 49 and representing persons age less 18 years. With this variable we align all the points on a common age based scale. The histogram of numbers of observations obtained for different t is given in appendix (chart 2). Table 6 in the appendix presents some descriptive statistics for observed pension points P for each value of t, while the extent of censoring for three regimes is sketched in table 7. Histogram for actual uncensored earnings for the out–of–sample individuals in 2001 is plotted on chart 3.

Fortunately, already table 6 displays an upside-down U–shape in the means of earnings that justify quadratic formulation of the lifecycle model that follows.

3. Modeling

In this paper we consider several models in search for best model which would have to be good enough in predicting earnings in the late years of working history but remain as simple as possible. With these concerns in mind we consider three formulations with constant over time coefficients: regular linear regression, standard one–sided tobit model, and precise double–sided censored model (models A, B and C respectively).

Model A is formulated as a classical linear regression with two explanatory variables: time index to the power of one and to the power of two. With index k denoting a particular observation we have

$$P_k = a + b\,t_k + c\,t_k^2 + \varepsilon_k,\ I_k = P_k + 1. \qquad (4)$$

Thus, we practically do not distinguish income from pension point apart from a small linear transformation following from formulas (1)–(3) and clearly seen on Chart 1. Disturbances ε_k are assumed to be independent and identically normally distributed with zero mean and constant variance. We use ordinary least squares estimation procedure to obtain the values of three parameters: *a*, *b* and *c*.

Undoubtedly linear regression is unsuitable formulation for the data in question, still, we estimate this model for several reasons. First, it will make it possible to relate current analysis to previous results. Second, linear regression will serve a practical purpose as producing starting values for maximum likelihood maximization in the later models. Third, since predicting earnings for a few years at the end of working history is often an internal task in bigger projects, it will be useful to identify how big is the error from using ordinary regression on pension points instead of the appropriate tobit–type model.

Model B is formulated as a standard Tobit model with the same explanatory variables and latent variable equal to annual earnings minus one.

$$I_k - 1 = a + b\,t_k + c\,t_k^2 + \varepsilon_k,\ P_k = \begin{cases} I_k - 1; I_k - 1 > 0, \\ 0; I_k - 1 \le 0. \end{cases} \tag{5}$$

Intuition for this formulation follows from Table 7 in the appendix. It seems that censoring from below plays a more important role in the data while high earnings are much more rare. Also in many applications it may be more important to be able to separate different low incomes other than to accurately predict large ones. Tobit model is estimated with standard statistical procedures.

Model C is developed as the most accurate model for incomes with pension points data. It utilizes formulas (1)–(3) to full extent and allows for several modifications with different assumptions about the distribution of disturbances. Our intention here is to model not only truncation from above and below but also changes in censoring regimes throughout the observation period.

First assume latent annual earnings I_k to be linear in parameters and quadratic in time structure.

$$I_k = a + b\,t_k + c\,t_k^2 + \varepsilon_k = A_k + \varepsilon_k. \tag{6}$$

Note that compared to the first two models parameter *a* now increased by one unit, so care should be taken when comparing the estimations. Assume ε_k are continuous independent

identically distributed random variables with cumulative distribution function $F(x)$ and density function $f(x)$. Introduce the following 11 index sets:

N_{11} – all observations in regime 1 with $P_k = 0$ ($I_k \le 1$).

N_{12} – all observations in regime 1 with $P_k \in (0;7)$. Then for $k \in N_{12}$ $I_k = P_k + 1 \in (1;8]$.

N_{13} – all observations in regime 1 with $P_k = 7$ ($I_k > 8$).

N_{21} – all observations in regime 2 with $P_k = 0$ ($I_k \le 1$).

N_{22} – all observations in regime 2 with $P_k \in (0;7]$. Then for $k \in N_{22}$ $I_k = P_k + 1 \in (1;8]$.

N_{23} – all observations in regime 2 with $P_k \in (7;\frac{25}{3})$. Then for $k \in N_{23}$ $I_k = 3P_k - 13 \in (8;12]$.

N_{24} – all observations in regime 2 with $P_k = \frac{25}{3}$ ($I_k > 12$).

N_{31} – all observations in regime 3 with $P_k = 0$ ($I_k \le 1$).

N_{32} – all observations in regime 3 with $P_k \in (0;5]$. Then for $k \in N_{32}$ $I_k = P_k + 1 \in (1;6]$.

N_{33} – all observations in regime 3 with $P_k \in (5;7)$. Then for $k \in N_{33}$ $I_k = 3P_k - 9 \in (6;12]$.

N_{34} – all observations in regime 3 with $P_k = 7$ ($I_k > 12$).

Defined sets $N_{11}..N_{34}$ repeat the case structure of formulas (1)–(3) and by construction compose a full system of mutually exclusive sets with respect to observation indexes. The inverse relations for I_k and P_k are given based on formulas (1)–(3).

Now it is possible to write down separate likelihood functions for the three regimes. We make use of the following result on conditional density function (here capital and small letters 'f' denote correspondingly cumulative distribution function and density function, $\tilde{x}$ stands for arbitrary random variable).

$$f_{\tilde{x}\in[\alpha;\beta]}(x \mid \tilde{x} \in [\alpha;\beta]) = \frac{f_{\tilde{x}}(x)}{F_{\tilde{x}}(\beta) - F_{\tilde{x}}(\alpha)}. \qquad (7)$$

We also put subscript of the corresponding random variable for its density function to avoid confusion. For regime 1 the log–likelihood looks as follows.

$$lnLF_1 = \sum_{k\in N_{11}} \ln\left[\Pr\{P_k = 0\}\right] + \sum_{k\in N_{12}} \ln\left[f^P\left(P_k \mid P_k \in (0,7]\right)\Pr\{P_k \in (0,7]\}\right] +$$
$$+ \sum_{k\in N_{13}} \ln\left[\Pr\{P_k = 7\}\right] = \sum_{k\in N_{11}} \ln\left[\Pr\{\varepsilon_k \le 1 - A_k\}\right] +$$

$$+\sum_{k\in N_{12}}\ln\left[f\left(P_k+1-A_k\mid\varepsilon_k\in(1-A_k,8-A_k]\right)\Pr\{\varepsilon_k\in(1-A_k,8-A_k]\}\right]+$$
$$+\sum_{k\in N_{13}}\ln\left[\Pr\{\varepsilon_k\geq 8-A_k\}\right]=\sum_{k\in N_{11}}\ln\left[F(1-A_k)\right]+\sum_{k\in N_{12}}\ln\left[f\left(P_k+1-A_k\right)\right]+$$
$$+\sum_{k\in N_{13}}\ln\left[1-F(8-A_k)\right] \qquad (8)$$

Likelihood function (8) follows directly from the mixture of discrete and continuous probability distributions for P_k with probability mass concentrated in points 0 and 7. Although departing from standard double–sided Tobit model [Maddala, 1988] of the first regime (in a sense that more intervals are introduces and scaling is done in addition to censoring) similar argument gives for regimes 2 and 3 consequently

$$lnLF_2=\sum_{k\in N_{21}}\ln\left[\Pr\{P_k=0\}\right]+\sum_{k\in N_{22}}\ln\left[f^P\left(P_k\mid P_k\in(0,7]\right)\Pr\{P_k\in(0,7]\}\right]+$$
$$+\sum_{k\in N_{23}}\ln\left[f^P\left(P_k\mid P_k\in(7,\frac{25}{3}]\right)\Pr\{P_k\in(7,\frac{25}{3}]\}\right]+\sum_{k\in N_{24}}\ln\left[\Pr\{P_k=\frac{25}{3}\}\right]=$$
$$=\sum_{k\in N_{21}}\ln\left[\Pr\{\varepsilon_k\leq 1-A_k\}\right]+$$
$$+\sum_{k\in N_{22}}\ln\left[f\left(P_k+1-A_k\mid\varepsilon_k\in(1-A_k,8-A_k]\right)\Pr\{\varepsilon_k\in(1-A_k,8-A_k]\}\right]+$$
$$+\sum_{k\in N_{23}}\ln\left[f\left(3P_k-13-A_k\mid\varepsilon_k\in(8-A_k,12-A_k]\right)\Pr\{\varepsilon_k\in(8-A_k,12-A_k]\}\right]+$$
$$+\sum_{k\in N_{24}}\ln\left[\Pr\{\varepsilon_k\geq 12-A_k\}\right]=\sum_{k\in N_{21}}\ln\left[F(1-A_k)\right]+\sum_{k\in N_{22}}\ln\left[f\left(P_k+1-A_k\right)\right]+$$
$$+\sum_{k\in N_{23}}\ln\left[f\left(3P_k-13-A_k\right)\right]+\sum_{k\in N_{24}}\ln\left[1-F(12-A_k)\right]. \qquad (9)$$

$$lnLF_3=\sum_{k\in N_{31}}\ln\left[\Pr\{P_k=0\}\right]+\sum_{k\in N_{32}}\ln\left[f^P\left(P_k\mid P_k\in(0,5]\right)\Pr\{P_k\in(0,5]\}\right]+$$
$$+\sum_{k\in N_{33}}\ln\left[f^P\left(P_k\mid P_k\in(5,7]\right)\Pr\{P_k\in(5,7]\}\right]+\sum_{k\in N_{34}}\ln\left[\Pr\{P_k=7\}\right]=$$
$$=\sum_{k\in N_{31}}\ln\left[\Pr\{\varepsilon_k\leq 1-A_k\}\right]+$$
$$+\sum_{k\in N_{32}}\ln\left[f\left(P_k+1-A_k\mid\varepsilon_k\in(1-A_k,6-A_k]\right)\Pr\{\varepsilon_k\in(1-A_k,6-A_k]\}\right]+$$
$$+\sum_{k\in N_{33}}\ln\left[f\left(3P_k-9-A_k\mid\varepsilon_k\in(6-A_k,12-A_k]\right)\Pr\{\varepsilon_k\in(6-A_k,12-A_k]\}\right]+$$
$$+\sum_{k\in N_{34}}\ln\left[\Pr\{\varepsilon_k\geq 12-A_k\}\right]=\sum_{k\in N_{31}}\ln\left[F(1-A_k)\right]+\sum_{k\in N_{32}}\ln\left[f\left(P_k+1-A_k\right)\right]+$$
$$+\sum_{k\in N_{33}}\ln\left[f\left(3P_k-9-A_k\right)\right]+\sum_{k\in N_{34}}\ln\left[1-F(12-A_k)\right]. \qquad (10)$$

Obviously complete log–likelihood function is simply the sum of the likelihood functions in three regimes.

$$lnLF = lnLF_1 + lnLF_2 + lnLF_3. \quad (11)$$

Maximizing *lnLF* with respect to structural parameters *a*, *b* and *c*[3] and also with respect to parameters of distribution leads to their ML estimates which inherit all the nice asymptotic properties.

As in Tobit model we assume normal distribution of disturbances, thus putting simply $F(x) = \Phi(\frac{x}{\sigma})$ and $f(x) = \frac{1}{\sigma}\phi(\frac{x}{\sigma})$ where $\Phi(x)$ and $\phi(x)$ are respectively standard normal cumulative distribution and density functions.

4. Estimation

Estimation of the models was carried out with statistical package TSP 4.5. For the first two models standard preprogrammed procedures were used while model C was estimated using general optimization procedure with manually programmed likelihood function. The results for the estimations are presented in Table 1.

Table 1. Estimation results.

Model	Parameter	Description	Estimator	St.err.	t-test	p-value
A	*a*	Constant term	0,22717	0,00241670	94,0	[.000]
	b	Time index	0,29011	0,00021694	1337,3	[.000]
	c	Squared time index	-0,00528	0,00000439	-1202,8	[.000]
B	*a*	Constant term	-0,54423	0,00279951	-194,4	[.000]
	b	Time index	0,34439	0,00024970	1379,24	[.000]
	c	Squared time index	-0,00627	0,00000504	-1243	[.000]
	σ	St.err. for residuals	2,39404	0,00044764	5348,18	[.000]
C	*a*	Constant term	0,04984	0,00366554	13,6	[.000]
	b	Time index	0,38236	0,00031722	1 205,4	[.000]
	c	Squared time index	-0,00670	0,00000616	-1 087,4	[.000]
	σ	St.err. for residuals	2,90066	0,00048264	6 010,0	[.000]

As we can see coefficients in all models were very sharply estimated and all present the same general pattern of flat concave parabola in lifetime earnings[4]. It's worth noting though that

[3] Given that the parameters are constant in time and are "regime invariant".

quite expectedly model A makes this parabola flatter than the censored models since boundary observations are treated as they were free. For the last two models standard error of the residuals is estimated along with structural parameters.

Compared to previous result in [Iskhakov, Kalvarskaia 2003] model A presents a steeper profile of earnings history with higher starting level. However, the difference is quite small and the results indeed appear to be well corresponding.

Table 2 presents the story told by the models by some reference points. Figures are very similar in the placing the life cycle earnings curve along the age line, but differ somewhat in positioning it over the incomes levels. Models A and B correspond very well in maximum average lifetime earnings, but differ in starting salary (showing again that linear regression gives flatter profile). Models A and C are similar in starting level, but most accurate model gives much higher maximum salary. Still, the figures in Tables 1 and 2 do not seem to vary very much, so the models seem to be very similar.

Table 2. Average picture of earnings throughout lifetime.

Model	Earnings when starting to work (age 19), in G	Maximum earnings during lifetime, in G	Age when maximum earnings are attained	Earnings when retiring (age 67), in G	Length of potential working life (age when earnings become 1G)
Model A	1,51	5,21	45	2,77	73
Model B	0,79	5,19	45	2,29	71
Model C	1,43	6,51	46	3,70	75

5. Evaluation and comparison

Table 1 purposely contains no measures of fit for the models. Current section of the paper specifically addresses the task of judging and comparing the models.

Since the data used to estimate the models was obtained from different individuals it is very much scattered on the vertical axes. This makes it rather unreasonable to measure goodness of fit in a standard way: all such measures would be quite poor because it is naturally

[4] Strictly speaking, interpretation of the parameters is different for each model, but for the sake of simplicity we use ”weak” life-cycle earnings parabola interpretation which is also plausible.

impossible to explain individual heterogeneity in the intercept term when looking for common shape of the equation. Best way to approach this task would be through introducing individual specific intercepts by means of dummy variables or through random effects. The former is very hard to perform with the number of individuals we are working with, the latter is left for another paper. Instead, we introduce individual specific intercept terms only after estimating the model when testing it and making predictions. This rather harsh way at first glance seems to be a bit simpler than the other two.

Our intention is to use estimations $\bar{b}$ and $\bar{c}$ of the shape parameters *b* and *c* for the lifetime earnings parabola and plug in individual specific intercept terms. Recalculation of the intercepts is based on comparison of the means of actually observed pension points for each person and the means of their predicted values – these two should be equal. An open question is whether all or some specific observations should be used for this calibration. We apply two possible approaches: equalize overall lifetime averages and use only three observations available before the year of prediction to apparently increase prediction accuracy.

For model A the estimated individual specific intercept term $\bar{a}_i$ is calculated according to simple formula (where *S* is the set of observations used for individual specific intercept calibration for particular person *i* and $|S| \leq 49$ denotes the power of set *S*).

$$\bar{a}_i = \frac{1}{|S|}\sum_{j \in S}\left(P_{ij} + 1 - \bar{b}t_j - \bar{c}t_j^2\right). \tag{12}$$

This formula assumes uncensored linear relationship between pension points and incomes and therefore can not be applied for other models. In the latter case we are looking for $\bar{a}_i$ that satisfies

$$\frac{1}{|S|}\sum_{j \in S}\mathrm{E}\left(P_{ij}(a_i + \bar{b}t_j + \bar{c}t_j^2) \mid \text{censoring}\right) = \frac{1}{|S|}\sum_{j \in S}P_{ij}, \tag{13}$$

that is the average of expected values of pension points as functions of predicted incomes subject to censoring should equal the average of observed pension points. Formula (13) takes particular forms for models B and C. For standard tobit model we use standard result on the expectation of the dependent variable (expectation in the sense of all values – censored and uncensored) to get

$$\frac{1}{|S|}\sum_{j \in S}\left[(\bar{a}_i + \bar{b}t_j + \bar{c}t_j^2)\Phi\left(\frac{\bar{a}_i + \bar{b}t_j + \bar{c}t_j^2}{\sigma}\right) + \sigma\phi\left(\frac{\bar{a}_i + \bar{b}t_j + \bar{c}t_j^2}{\sigma}\right)\right] = \frac{1}{|S|}\sum_{j \in S}(P_{ij} + 1). \tag{14}$$

Solving (14) for $\bar{a}_i$ proves difficult, we use an approximate solution of the form

$$\bar{a}_i = \frac{\sum_{j\in S}\left(P_{ij} + 1 - \Phi\left(\frac{\bar{a}+\bar{b}t_j+\bar{c}t_j^2}{\sigma}\right)(\bar{b}t_j+\bar{c}t_j^2) - \sigma\phi\left(\frac{\bar{a}+\bar{b}t_j+\bar{c}t_j^2}{\sigma}\right)\right)}{\sum_{j\in S}\Phi\left(\frac{\bar{a}+\bar{b}t_j+\bar{c}t_j^2}{\sigma}\right)} \qquad (15)$$

by substituting individual specific coefficient $\bar{a}_i$ with average over whole sample $\bar{a}$ when calculating censoring effects.

In the last model expression for $\bar{a}_i$ is so nasty that it deserves to be removed to appendix.

We compare the models by the accuracy of out–of–sample predictions. As mentioned in the first part 1000 observations of persons turning 60 in the year 2001 were laid aside to compose separate testing space for the models. Predictions of uncensored annual earnings obtained with individual intercepts are compared to actually observed in year 2001 incomes. Table 3 contains several different sums of squares of prediction deviations for all models.

Table 3. Comparing models with sums of squares of deviations SSD (* row minimums).

	Model A	Model B	Model C
Total sum of squared deviations	9553,64	9590,69	9453,70*
SSD when prediction is higher than observation	3294,76	3126,65*	3785,94
SSD when prediction is lower than observation	6258,88	6464,04	5667,76*
SSD for high observed incomes ($I \geq 5G$)	5822,47	5880,85	5041,84*
SSD for low observed incomes ($I < 2G$)	3056,60	2889,14*	3360,11

We use several criteria to judge accuracy of the models. The main one measuring overall accuracy is sum of squares of differences between observed and predicted earnings. Table 3 clearly shows that the most accurate model C is outperforming the first two according to this means of comparison although the difference does not seem to be too drastic.

In addition to the general criterion four minor ones allow for more detailed comparison of the models. Separate sums of squares of deviations are calculated for over– and underestimating, and also for high and low observed incomes (respectively higher than 5G and lower than 2G). The picture is the following. Models B and C perform better than linear regression in all cases, but their behavior is quite different. Model C tend to overestimate earnings while model B tend to underestimate them (this follows from the fact that model B has less upwards deviations while model C – downwards, Table 3). But in doing so model B performs

considerably better on low incomes with model C doing much better on high incomes. This coincides well with results from Table 2. Taking into account only censoring from below implies that model B positions the lifetime earnings curve much lower also making it a bit flatter compared to model C. Thus, the latter grasps high incomes much better letting model B to perform better on low incomes. One solid consequence from comparison of Table 3 is the worst performance by linear regression in all considered cases, making this model truly unfavorable for lifetime incomes prediction.

Table 4. Absolute errors in predictions (* minimums among models).

		N	Mean	Std	Max
Absolute errors	Model A	1000	2,22	2,15	24,52
	Model B	1000	2,23	2,15	24,61
	Model C	1000	2,17*	2,18	24,29*
Deviations up from observed values	Model A	385	2,42	1,65	7,34*
	Model B	371	2,37	1,68	7,51
	Model C	427	2,29*	1,91	8,54
Deviations down from observed values	Model A	615	2,10	2,41	24,52
	Model B	629	2,15	2,38	24,61
	Model C	573	2,08*	2,36	24,29*

Table 4 presents a slightly different approach for model comparison which substantiate the results already obtained. Here we can analyze absolute values of the deviations of predictions from observed values instead of their sums of squares. Thus, we are able to look at mean errors, max errors and standard deviations of errors made by the models. Again when comparing mean total errors model C outperforms the other two. But the difference is not that high – all models on average are off by 2,17–2,23G anyway. Interesting that even though model C has highest variance in the errors it remains the best model from the point of view of min–max criterion. When looking at positive and negative deviations surprisingly linear regression becomes the model with smallest maximum error, but comparing means leads to model C dominance in all the cases.

Quite worrying fact that follows from table 4 is that all models are substantially off the target in predictions, especially in relative terms. Average predicted earnings at age 60 (calculated from the estimates, see table 1) are 4,10G, 3,87G and 5,29G correspondingly for models A, B and C. So, average prediction errors constitute correspondingly 54%, 58% and 41% of the

average predictions, that is of course unsatisfactory. Partial explanation for this poor performance follows from the simplicity of the models that are tested on quite complex prediction task. In the same time prediction accuracy can be easily increased by far with a different procedure for calculating individual specific intercepts. To illustrate this we include in the set *S* in formulas (12) and (13) only three last available observations of pension points for each individual. Thus, the earnings curve is set so that it perfectly matches the average of last three observations before the prediction year.

With the new predictions calculated with the new individual intercepts tables 3 and 4 should be recalculated and reconsidered. However, the relative performance by the models does not change, so all statements concerning comparison of the models are in effect. Table of sums of squares of deviations fully repeats the structure of table 3 and is therefore omitted. We include table 5 which presents absolute errors for new predictions.

Table 5. Absolute errors in predictions II (* minimums among models).

		N	Mean	Std	Max
Absolute errors	Model A	1000	1,27	1,97	22,98
	Model B	1000	1,17	1,95	22,85
	Model C	1000	1,15*	1,82	21,96*
Deviations up from observed values	Model A	426	1,24	1,24	7,12*
	Model B	420	1,10*	1,31	7,40
	Model C	524	1,13	1,35	8,89
Deviations down from observed values	Model A	574	1,28	2,37	22,98
	Model B	580	1,22	2,31	22,85
	Model C	476	1,17*	2,22	21,96*

It follows from table 5 that absolute prediction errors with the recalculated individual intercepts are reduced almost by half, but the comparative analysis leads to practically identical results. Again, model C takes all nominations apart from maximum positive deviation, in which standard linear regression appears to be best as before, and minimum mean positive deviations, where this time model B outperforms the other two. Still, the differences among the models in terms of size of absolute errors don't seem to be drastic.

Overall, in most cases model C performs best and whenever possible should be preferred to the other two, but the difference in predicting quality is quite small and it is quite possible to get away with the simpler model B.

5. Conclusion

The paper was devoted to the appropriate choice of a model to solve the problem of earnings prediction with the use of data on accumulated pension rights in terms of 'pension points'. The task of such predictions is inevitable in many labour market discrete choice studies since data at hand usually represents only the characteristics of the alternative chosen by the agent and no other. Norwegian statistics authorities have kept a beautiful record of earnings histories nearly for whole population which makes the prediction problem to be perceived as easily solvable. But the lifetime earnings data is kept in censored variables with changing over time censoring regimes. This makes predicting task more challenging requiring special Tobit–type models to be developed. Since earnings prediction is just an auxiliary task, developing a full scale model may be too costly. In this circumstances a reasonable question of finding the right balance between model complexity and the accuracy of predictions comes into play.

In the current paper three possible models for lifetime earnings were estimated and tested for out–of– sample prediction accuracy in order to address the described question of balance. Our main findings are the following.

As expected, the fine model with double–sided censoring and regime changes performed best overall, simple linear regression with no attention to censoring was the worst one with standard one–side censored Tobit model being intermediate in prediction quality. On the other hand average absolute errors in predictions made by the models are of very similar scale which justifies the use of even linear regressions in earnings predictions at least at the last stages of working life (our test was done for age 60). Additionally it can be noted that one–sided standard Tobit model is preferable when research is dealing a lot with low earnings, but in order to asses high income portion of the population the finest model is definitely necessary.

References

1. Greene, William H. Econometric Analysis. New Jersey, 2000.
2. Haugen, Frederik Insentivvirkninger av skatte-og pensjonsregel // Working paper, Frisch Center for Economic Research 4/2000, 2000.
3. Iskhakov, Fedor & Kalvarskaia, Maria AFP and OP data construction techniques // Working paper, Frisch Center for Economic Research 1/2003. Oslo, 2003.
4. Maddala, G.S. Limited dependent and qualitative variables in econometrics // Econometric Society Monographs No. 3. Cambridge New York New Rochelle Melbourne Sydney, 1988.
5. Røgeberg, Ole J. Married man and early retirement under AFP scheme // Memorandum of The Department of Economics, University of Oslo 02/2000, 2000.
6. Sydsæter, Knut & Strøm, Arne & Berck, Peter Economists' mathematical manual. Oslo, 2000.

Appendix

Table 6. Summary of distributions of pension points conditional on time index.

t	Observations	Min	Q1	Mean	Q3	Max	Std
2	51 416	0,00	0,00	0,84	1,50	7,00	1,07
3	105 060	0,00	0,00	1,03	1,85	7,00	1,19
4	153 274	0,00	0,00	1,39	2,41	7,00	1,38
5	199 596	0,00	0,00	1,68	2,85	7,00	1,54
6	240 487	0,00	0,00	1,92	3,19	8,33	1,68
7	277 841	0,00	0,00	2,13	3,46	8,33	1,79
8	309 065	0,00	0,00	2,34	3,72	8,33	1,90
9	342 228	0,00	0,03	2,52	3,95	8,33	2,01
10	374 191	0,00	0,12	2,68	4,19	8,33	2,10
11	404 566	0,00	0,24	2,83	4,40	8,33	2,19
12	433 161	0,00	0,41	2,96	4,58	8,33	2,26
13	460 326	0,00	0,58	3,08	4,74	8,33	2,31
14	485 738	0,00	0,77	3,18	4,86	8,33	2,34
15	510 911	0,00	0,96	3,27	4,96	8,33	2,36
16	510 911	0,00	1,18	3,37	5,06	8,33	2,37
17	510 911	0,00	1,38	3,48	5,15	8,33	2,36
18	510 911	0,00	1,57	3,58	5,24	8,33	2,35
19	510 911	0,00	1,75	3,68	5,33	8,33	2,34
20	510 911	0,00	1,92	3,79	5,41	8,33	2,32
21	510 911	0,00	2,10	3,89	5,50	8,33	2,29
22	510 911	0,00	2,28	4,00	5,61	8,33	2,27
23	510 911	0,00	2,44	4,10	5,70	8,33	2,25
24	510 911	0,00	2,57	4,17	5,77	8,33	2,22
25	510 911	0,00	2,67	4,23	5,81	8,33	2,20
26	510 911	0,00	2,75	4,28	5,84	8,33	2,19
27	510 911	0,00	2,82	4,28	5,75	8,33	2,14
28	510 909	0,00	2,85	4,26	5,67	8,33	2,10
29	510 907	0,00	2,87	4,24	5,61	8,33	2,07
30	510 905	0,00	2,89	4,22	5,55	8,33	2,04

t	Observations	Min	Q1	Mean	Q3	Max	Std
31	510 903	0,00	2,88	4,19	5,50	8,33	2,02
32	510 903	0,00	2,87	4,15	5,46	8,33	2,01
33	510 903	0,00	2,83	4,12	5,43	8,33	2,01
34	510 902	0,00	2,77	4,06	5,40	8,33	2,01
35	510 901	0,00	2,67	3,99	5,36	8,33	2,02
36	459 485	0,00	2,55	3,92	5,32	8,33	2,04
37	405 843	0,00	2,40	3,83	5,29	8,33	2,05
38	357 631	0,00	2,23	3,73	5,25	8,33	2,07
39	311 308	0,00	2,02	3,61	5,19	8,33	2,08
40	270 419	0,00	1,76	3,45	5,14	7,00	2,09
41	233 065	0,00	1,48	3,33	5,10	7,00	2,13
42	201 841	0,00	1,08	3,17	5,05	7,00	2,17
43	168 679	0,00	0,41	2,94	4,98	7,00	2,21
44	136 717	0,00	0,07	2,61	4,57	7,00	2,20
45	106 342	0,00	0,00	2,10	4,00	7,00	2,19
46	77 747	0,00	0,00	1,68	3,27	7,00	2,09
47	50 582	0,00	0,00	1,31	2,36	7,00	1,94
48	25 173	0,00	0,00	1,00	1,26	7,00	1,76

Table 7. Extent of censoring in different regimes, number of observations.

	Censoring regime					
	1960-1970		1971-1991		1992-2000	
Censored from above	47 487	2,32%	236 208	2,20%	180 780	3,93%
Not censored	1 402 780	68,64%	9 477 031	88,33%	3 910 402	85,04%
Censored from below	593 377	29,04%	1 015 892	9,47%	506 900	11,02%
Total	2 043 644	100,00%	10 729 131	100,00%	4 598 082	100,00%

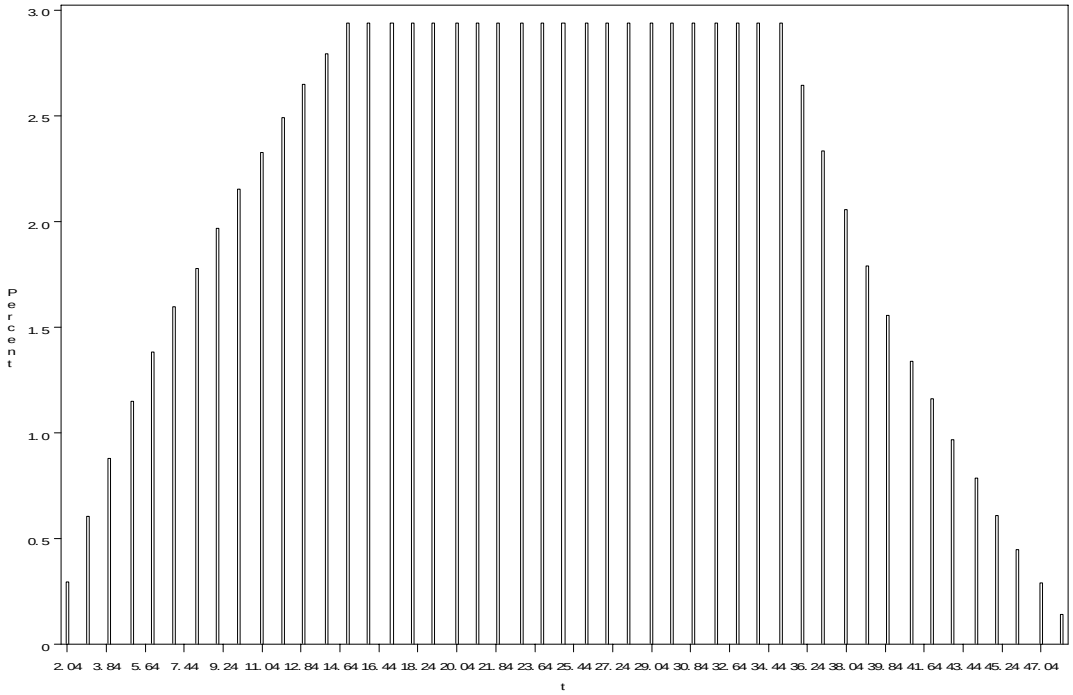

Chart 2. Relative number of observations for each value of *t*.

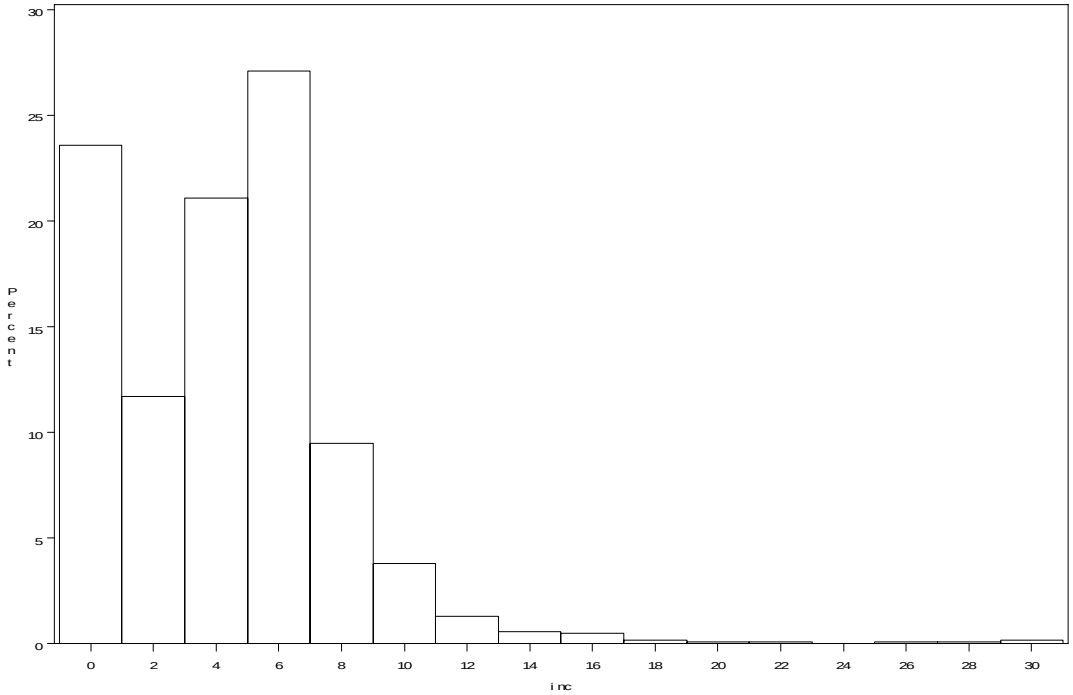

Chart 3. Distribution of uncensored incomes for 1000 individuals “out of sample”.

Deriving individual specific intercept coefficient for model C

Let S_1, S_2 and S_3 correspond to the years of observation under corresponding regimes, $S_1 \cap S_2 \cap S_3 = S$. Then for three regimes write (with notation introduced in (6) where estimated values take place of the corresponding coefficients)

$$\mathrm{E}_1(P) = \Pr\{I<1\}\mathrm{E}(P|P=0) + \Pr\{I\in(1,8)\}\mathrm{E}(I-1|I\in(1,8)) + \Pr\{I>8\}\mathrm{E}(P|P=7) =$$

$$= \Pr\{\varepsilon\in(1-A,8-A)\}\mathrm{E}(A+\varepsilon-1|\varepsilon\in(1-A,8-A)) + 7\Pr\{\varepsilon>8-A\} =$$

$$= \left[F(8-A)-F(1-A)\right]\cdot\left[A-1+\int_{1-A}^{8-A} x\frac{f(x)}{F(8-A)-F(1-A)}dx\right] + 7(1-F(8-A)) =$$

$$= \left[F(8-A)-F(1-A)\right](A-1)+\int_{1-A}^{8-A} xf(x)dx \; + 7(1-F(8-A)). \qquad (16)$$

$$\mathrm{E}_2(P) = \Pr\{I<1\}\mathrm{E}(P|P=0) + \Pr\{I\in(1,8)\}\mathrm{E}(I-1|\, I\in(1,8]) +$$

$$\Pr\{I\in(8,12)\}\mathrm{E}\left(\frac{13}{3}+\frac{I}{3}\,|I\in(8,12)\right) + \Pr\{I>12\}\mathrm{E}\left(P|P=\frac{25}{3}\right) = \Pr\{\varepsilon\in(1-A,8-A)\}\mathrm{E}(A+\varepsilon-1|\varepsilon\in(1-A,8-A)) +$$

$$+ \Pr\{\varepsilon\in(8-A,12-A)\}\mathrm{E}\left(\frac{13}{3}+\frac{A+\varepsilon}{3}\,|\varepsilon\in(8-A,12-A)\right) + \frac{25}{3}\Pr\{\varepsilon>12-A\} =$$

$$= \left[F(8-A)-F(1-A)\right]\cdot\left[A-1+\int_{1-A}^{8-A} x\frac{f(x)}{F(8-A)-F(1-A)}dx\right] +$$

$$+ \left[F(12-A)-F(8-A)\right]\cdot\left[\frac{13+A}{3}+\frac{1}{3}\int_{8-A}^{12-A} x\frac{f(x)}{F(12-A)-F(8-A)}dx\right] + \frac{25}{3}(1-F(12-A)) =$$

$$= \left[F(8-A)-F(1-A)\right](A-1)+\int_{1-A}^{8-A} xf(x)dx \; +$$

$$+ \left[F(12-A)-F(8-A)\right]\frac{13+A}{3}+\frac{1}{3}\int_{8-A}^{12-A} xf(x)dx \; + \frac{25}{3}(1-F(12-A)). \qquad (17)$$

$$\mathrm{E}_3(P) = \Pr\{I<1\}\mathrm{E}(P|P=0) + \Pr\{I\in(1,6)\}\mathrm{E}(I-1|\, I\in(1,6]) +$$

$$\Pr\{I\in(6,12)\}\mathrm{E}\left(3+\frac{I}{3}|I\in(6,12)\right) + \Pr\{I>12\}\mathrm{E}(P|P=7) = \Pr\{\varepsilon\in(1-A,6-A)\}\mathrm{E}(A+\varepsilon-1|\varepsilon\in(1-A,6-A)) +$$

$$+ \Pr\{\varepsilon\in(6-A,12-A)\}\mathrm{E}\left(3+\frac{A+\varepsilon}{3}\,|\varepsilon\in(6-A,12-A)\right) + 7\Pr\{\varepsilon>12-A\} =$$

$$= \left[F(6-A)-F(1-A)\right]\cdot\left[A-1+\int_{1-A}^{6-A} x\frac{f(x)}{F(6-A)-F(1-A)}dx\right] +$$

$$+\left[F(12-A)-F(6-A)\right]\cdot\left[3+\frac{A}{3}+\frac{1}{3}\int_{6-A}^{12-A}x\frac{f(x)}{F(12-A)-F(6-A)}dx\right]+7(1-F(12\text{-}A))=$$

$$=\left[F(6-A)-F(1-A)\right](A-1)+\int_{1-A}^{6-A}xf(x)dx\ +$$

$$+\left[F(12-A)-F(6-A)\right]\frac{9+A}{3}+\frac{1}{3}\int_{6-A}^{12-A}xf(x)dx\ +7(1-F(12\text{-}A)). \qquad (18)$$

When considering three regims expression (13) modifies to

$$\frac{1}{|S|}\left[\sum_{j\in S_1}\mathrm{E}_1\left(P_{ij}\right)+\sum_{j\in S_2}\mathrm{E}_2\left(P_{ij}\right)+\sum_{j\in S_3}\mathrm{E}_3\left(P_{ij}\right)\right]=\frac{1}{|S|}\sum_{j\in S}P_{ij}\,. \qquad (19)$$

Again, equation (19) proves to be impossible to solve analitically, I calculate censoring effects using sample average value of coefficient *a*. Moreover, now I have additional integrals with complex limits which must be calculated. Fortunately, with normal distribution of error terms it is not hard.

$$\int_{\alpha}^{\beta}xf(x)dx=\frac{1}{\sqrt{2\pi}}\int_{\alpha}^{\beta}\frac{x}{\sigma}\exp(-\frac{x^2}{2\sigma^2})dx=-\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2\sigma^2})\Big|_{\alpha}^{\beta}=\sigma\left[f(\alpha)-f(\beta)\right]. \qquad (20)$$

Finally, plugging (16)-(18) into (19) and rearranging, arrive at final result (recall that A = $\bar{a}\ +\ \bar{b}\ t_j\ +\ \bar{c}\ t_j^2$ and see next page).

$$a_i = \frac{\sum_{j\in S} P_{ij} - \sum_{j\in S_1}\left\{\left[\Phi\left(\frac{8-A}{\sigma}\right)-\Phi\left(\frac{1-A}{\sigma}\right)\right]\left[\bar{b}t_j+\bar{c}t_j^2-1\right]+\sigma\left[f\left(\frac{1-A}{\sigma}\right)-f\left(\frac{8-A}{\sigma}\right)\right]+7\left[1-\Phi\left(\frac{8-A}{\sigma}\right)\right]\right\}}{\sum_{j\in S_1}\left[\Phi\left(\frac{8-A}{\sigma}\right)-\Phi\left(\frac{1-A}{\sigma}\right)\right]+\sum_{j\in S_2}\left[\frac{1}{3}\Phi\left(\frac{12-A}{\sigma}\right)+\frac{2}{3}\Phi\left(\frac{8-A}{\sigma}\right)-\Phi\left(\frac{1-A}{\sigma}\right)\right]+\sum_{j\in S_3}\left[\frac{1}{3}\Phi\left(\frac{12-A}{\sigma}\right)+\frac{2}{3}\Phi\left(\frac{6-A}{\sigma}\right)-\Phi\left(\frac{1-A}{\sigma}\right)\right]} -$$

$$-\frac{\sum_{j\in S_2}\left\{\left[\Phi\left(\frac{8-A}{\sigma}\right)-\Phi\left(\frac{1-A}{\sigma}\right)\right]\left[\bar{b}t_j+\bar{c}t_j^2-1\right]+\left[\Phi\left(\frac{12-A}{\sigma}\right)-\Phi\left(\frac{8-A}{\sigma}\right)\right]\frac{13+\bar{b}t_j+\bar{c}t_j^2}{3}+\sigma\left[f\left(\frac{1-A}{\sigma}\right)-\frac{2}{3}f\left(\frac{8-A}{\sigma}\right)-\frac{1}{3}f\left(\frac{12-A}{\sigma}\right)\right]+\frac{25}{3}\left[1-\Phi\left(\frac{12-A}{\sigma}\right)\right]\right\}}{\bullet\bullet\bullet} -$$

$$-\frac{\sum_{j\in S_3}\left\{\left[\Phi\left(\frac{6-A}{\sigma}\right)-\Phi\left(\frac{1-A}{\sigma}\right)\right]\left[\bar{b}t_j+\bar{c}t_j^2-1\right]+\left[\Phi\left(\frac{12-A}{\sigma}\right)-\Phi\left(\frac{6-A}{\sigma}\right)\right]\frac{9+\bar{b}t_j+\bar{c}t_j^2}{3}+\sigma\left[f\left(\frac{1-A}{\sigma}\right)-\frac{2}{3}f\left(\frac{6-A}{\sigma}\right)-\frac{1}{3}f\left(\frac{12-A}{\sigma}\right)\right]+7\left[1-\Phi\left(\frac{12-A}{\sigma}\right)\right]\right\}}{\bullet\bullet\bullet}. \quad (21)$$

Frisch Centre Publications

All publications are available in Pdf-format at : www.frisch.uio.no

Rapporter (Reports)

1/2004	**Causality and Selection in Labour Market Transitions. Dissertation for the Dr.Polit Degree**	Tao Zhang
2/2004	**Arbeidstilbud når svart arbeid er en mulighet**	Tone Ognedal, Øystein Jørgensen, Steinar Strøm
3/2004	**Er det lengden det kommer an på? – Hvordan arbeidslediges jobbmuligheter påvirkes av nivået på dagpengene og hvor lenge de har gått ledig**	Øystein Jørgensen
4/2004	**Pris- og avanseregulering for legemidler**	Dag Morten Dalen, Steinar Strøm
5/2004	**Statlig styring av prosjektledelse**	Dag Morten Dalen, Ola Lædre, Christian Riis
6/2004	**Veier inn i, rundt i, og ut av det norske trygde- og sosialhjelpssystemet**	Elisabeth Fevang, Knut Røed, Lars Westlie, Tao Zhang
7/2004	**Undersysselsatte i Norge: Hvem, hvorfor og hvor lenge?**	Elisabeth Fevang, Knut Røed, Oddbjørn Raaum, Tao Zhang
8/2004	**Realopsjoner og fleksibilitet i store offentlige investeringsprosjekter**	Kjell Arne Brekke
9/2004	**Markeder med svart arbeid**	Erling Barth, Tone Ognedal
10/2004	**Skatteunndragelse og arbeidstilbud. En empirisk analyse av arbeidstilbudet når svart arbeid er en mulighet**	Kristine von Simson
1/2005	**Pliktige elsertifikater**	Rolf Golombek, Michael Hoel
2/2005	**En empirisk analyse av indeksprissystemet i det norske legemiddelmarkedet**	Tonje Haabeth

Arbeidsnotater (Working papers)

1/2004	**Samtidig bruk av Trygdeetaten, Arbeidsmarkedsetaten og Sosialtjenesten**	Morten Nordberg, Lars Westlie
2/2004	**Arbeidsledighet og svart arbeid. En empirisk**	Øyvind Johan Dahl

	analyse 1980 – 2003	
1/2005	**Lifetime earnings**	Fedor Iskhakov

Memoranda (Discussion papers)

The series is published by Department of Economics, University of Oslo, in co-operation with the Frisch Centre. This list includes memoranda related to Frisch Centre projects.
The complete list of memoranda can be found at www.sv.uio.no/sosoek/memo/.

1/2004	**To What Extent Is a Transition into Employment Associated with an Exit from Poverty?**	Taryn Ann Galloway
2/2004	**A dissolving paradox: Firms' compliance to environmental regulation**	Karine Nyborg, Kjetil Telle
4/2004	**Rainfall, Poverty and Crime in 19th Century Germany**	Halvor Mehlum, Edward Miguel, Ragnar Torvik
5/2004	**Climate policies and induced technological change: Impacts and timing of technology subsidies**	Snorre Kverndokk, Knut Einar Rosendahl, Thomas F. Rutherford
10/2004	**The sahdow economy in Norway: Demand for currency approach**	Isilda Shima
11/2004	**Climate Agreement and Technology Policy**	Rolf Golombek, Michael Hoel
12/2004	**The Norwegian market for pharmaceuticals and the non-mandatory substation reform of 2001: the case of enalapril**	Tiziano Razzolini
13/2004	**Sectoral labor supply, choice restrictions and functional form**	John K. Dagsvik, Steinar Strøm
17/2004	**Unilateral emission reductions when there are cross - country technology spillovers**	Rolf Golombek, Michael Hoel
25/2004	**Moral hazard and moral motivation: Corporate social responsibility as labor market screening**	Kjell Arne Brekke, Karine Nyborg
26/2004	**Can a carbon permit system reduce Spanish unemployment?**	T. Fæhn, A. G. Gómez-Plana, S. Kverndokk
5/2005	**The Kyoto agreement and Technology Spillovers**	Rolf Golombek, Michael Hoel
6/2005	**Labor supply when tax evasion is an option**	Øystein Jørgensen, Tone Ognedal, Steinar Strøm
9/2005	**The Fear of Exclusion: Individual Effort when Group Formation is Endogenous**	Kjell Arne Brekke, Karine Nyborg, Mari Rege
11/2005	**Tax evasion and labour supply in Norway in 2003: Structural models versus flexible functional form models**	Kari Due-Andresen

The Frisch Centre

The Ragnar Frisch Centre for Economic Research is an independent research institution founded by the University of Oslo. The Frisch Centre conducts economic research in co-operation with the Department of Economics, University of Oslo. The research projects are mostly financed by the Research Council of Norway, government ministries and international organisations. Most projects are co-operative work involving the Frisch Centre and researchers in other domestic and foreign institutions.

Ragnar Frisch Centre for Economic Research
Gaustadalléen 21
N-0349 Oslo, Norway
T + 47 22 95 88 10
F + 47 22 95 88 25
frisch@frisch.uio.no
www.frisch.uio.no