|  |  |
| --- | --- |
| **MSc Economics**  **Dissertation**  28/08/2023  Omar Achour  13818207 | Home — Birkbeck, University of London |

A Markov Decision Process in a game of study:

A discrete choice model for students’ investment decisions in higher education

**Abstract:**

In this paper, we introduce a model for prospective students’ decision to join higher education and the optimal quantity to undertake. We set up a baseline model using a Markov Decision Process and depart from a perfect information case to include scenarios of omitted information about the agent’s reward, its transition probabilities and friction costs.

**7937 words**

**Keywords**: Markov Decision Process, Information Theory, Stochastic model, Transition matrix, Reward.

**Software:** RStudio, R (programming language)

Table of Contents

[1. Introduction 4](#_Toc144119773)

[2. Literature review 5](#_Toc144119774)

[3. Data 7](#_Toc144119775)

[3.1. Dataset 7](#_Toc144119776)

[3.2. Derived data 9](#_Toc144119777)

[3.3. Descriptive statistics 10](#_Toc144119778)

[4. Model 11](#_Toc144119779)

[4.1. Markov Decision Process 11](#_Toc144119780)

[4.2. Model specification 12](#_Toc144119781)

[4.3. Reinforcement learning methods 12](#_Toc144119782)

[4.3.1. Policy iteration 12](#_Toc144119783)

[4.3.2. Value iteration 13](#_Toc144119784)

[4.3.3. Q-learning 13](#_Toc144119785)

[4.4. Stochastic model of rewards 13](#_Toc144119786)

[5. Findings 14](#_Toc144119787)

[5.1. Complete information case 14](#_Toc144119788)

[5.2. Incomplete information case 16](#_Toc144119789)

[5.2.1. Hidden transition probabilities 16](#_Toc144119790)

[5.2.2. Hidden rewards 17](#_Toc144119791)

[5.2.3. Cost evaluation 18](#_Toc144119792)

[5.2.4. Synthetic model 19](#_Toc144119793)

[5.2.5. Evidence of reversal in self-identification 20](#_Toc144119794)

[6. Synthesis 22](#_Toc144119795)

[7. Conclusion 22](#_Toc144119796)

[References 23](#_Toc144119797)

[Appendix 26](#_Toc144119798)

Table of Figures

[**Figure 1: College credits accumulation by year (left), College on income(right) 7**](#_Toc143435686)

[**Figure 2: College credits on mental ability (left) and self-esteem (right) 8**](#_Toc143435687)

[**Figure 3: Income in logs on experience (left) and college (right) in year 9**](#_Toc143435688)

[**Figure 4: Descriptive statistics table 10**](#_Toc143435689)

[**Figure 5: Policy evaluation of ‘pursuing further study’ 14**](#_Toc143435690)

[**Figure 6: Optimal policy and Bellman operator of ‘pursuing further study’ 15**](#_Toc143435691)

[**Figure 7: Policy iteration for patient agents 15**](#_Toc143435692)

[**Figure 8: Policy iteration of for less patient agents 16**](#_Toc143435693)

[**Figure 9: Q-learning optimal policy by discount factors 16**](#_Toc143435694)

[**Figure 10: Log-linear model of income on college, experience and ability 17**](#_Toc143435695)

[**Figure 11: Synthetic and original log-income rewards 18**](#_Toc143435696)

[**Figure 12: Value iteration for synthetic and original log-income rewards 18**](#_Toc143435697)

[**Figure 13: Cost function 19**](#_Toc143435698)

[**Figure 14: Synthetic model with omitted transition and reward matrices against costs 20**](#_Toc143435699)

[**Figure 15: Synthetic model with omitted rewards and frictional costs 20**](#_Toc143435700)

# 1. Introduction

Long has been the debate about education and its impact on earnings, if education is to be viewed simply as a signal, its utility is deprived of a component of self-fulfilment and learning primordial to each rational being. Nevertheless, an exclusively monetary approach to quantify the time, effort and very often monetary costs involved in such an investment would certainly provide and insight on the worthiness of this process.

Previous studies have been restrained to regression analyses to extract information. In this paper, we ought to introduce a novel approach as a discrete decision, one-player game using Markov decision processes. This would allow us to model the prospective student’s decisions with regards to higher education and the optimal quantity to undertake. It would help answer specific questions relating to how long should one study at university, up to which degree and when is it imperative to continue to the next level. These decisions are modelled in a state-action framework coupled with a reward function that traces yearly income and a transition matrix that determines the agent’s possible moves.

In answering the question of what the optimal duration of study is and whether characteristics of patience or individual’s type influence this decision and whether other factors might be more prominent in determining lifetime earnings, we aim to assess the *impact of information available to the agent in determining the optimal decision* and whether the absence of some elements would deter the agent to choose otherwise. To do so we set out a baseline model with full information and gradually omit core elements of the Markov decision process. In the baseline model, we evaluate an optimal policy that we solve for using an iterative method. We then are able to employ the Q-learning method, of reinforcement learning, to simulate a process without a transition matrix. Then build a stochastic model of income determined by education, work experience and an ability measure in the case of hidden rewards. We approach furthermore a realistic setting by introducing an evaluation of the underlying costs of education in public and private schools.

In the next section, we review the most prominent literature in job-market signalling, education, earnings and self-selection in an information theory setting. In section 3, we describe the data in hand, the variables selection and the data-mining process. In the following section, we layout the mathematical background of Markov decision processes, with some reinforcement learning methods and a stochastic model of rewards. In section 5, we discuss our findings at length and introduce the incomplete information approach which we review in the synthesis.

# 2. Literature review

In most job markets the employer when hiring, is uncertain about the productive capabilities of employees. The fact that it takes time to learn an individual's productive unknown beforehand means that it is an investment decision under uncertainty. Spence (1973) argues that to hire someone is akin to a gamble, the wage is taken to be the individual's marginal contribution to the hiring organization.

On the basis of previous experience however, the employer forms conditional probability expectations given certain signals, these are alterable at cost. The signalling costs are in the case of education tuition fees, psychological effort and time spent studying. Individuals, when acquiring an education, need not think of themselves as signalling but rather as an investment decision to be pursued if there is sufficient return as defined by the offered wage schedule. Individuals, then, are assumed to select signals so as to maximize the difference between offered wages and signalling costs.

Stiglitz (1975) discusses screening in education, the economic costs and benefits of labelling, the institutions that provide it, and the determination of the equilibrium amount of screening under various institutional arrangements. In his example, the individual is capturing his ability rents which in the absence of screening he shares with others. This approach is based on a set of assumption: The more able are better in every relevant sense than the less able, labour is inelastically supplied and there are no increases in production from sorting individuals. Furthermore, individuals have perfect information about their own abilities, and although no method of on-the-job screening is available, the screening beforehand is perfectly accurate with the information acquired considered fairly general.

There are various mechanisms to provide screening information and the argument is that if the education system does any sorting, the groups into which any individual has been selected gives some information to the firm about that individual, this can be the major a student chooses or the university he attends. Performance tests are another mechanism, where individuals are dealt with identical learning schedules but some learn better than others, which can be quantified by a grade.

A great deal of information is however the result of self-selection, this mechanism considers attributes of an agent about which he has more information than the firm. Some individuals have "more" of a given attribute than others and are thus willing to take the risk in order to show their ability. It should be noted that self- selection only works because of performance tests. If there were no chances of failure, everyone would attempt to go to the best school and everyone would pass from one grade to the other.

Rothchild and Stiglitz (1976) analyse a model of competitive insurance markets with imperfect information, in an approach that can be adapted to education or other signalling schemes. They show that if individuals were willing or able to reveal their information, everybody could be made better off and that high-risk individuals, simply by being in the cohort cause an externality, the low-risk individuals are worse off than they would be in the absence of the high-risk individuals. However, the high-risk individuals are no better off than they would be in the absence of the low-risk individuals.

Riley (1979) argues that if buyers are less knowledgeable about a product’s quality than sellers, market prices will reflect average quality. Sellers of high-quality products therefore have an incentive to engage in some distinguishing activity which serves as a signal to potential buyers. He explores the viability of such signalling and an alternative non-cooperative equilibrium concept is then developed in which potential price searching agents take account of possible reactions by other agents. It is shown that there is a unique reactive informational equilibrium.

Willis (1979) devises a structural model of the demand for college attendance as a combination of the theory of comparative advantage and statistical models of self-selection and unobserved ability. His estimates from NBER-Thorndike data strongly support the theory. Expected lifetime earnings gains influence college attendance decisions. Individuals who did not attend college would have earned less than similar people who did, while those who attended college would have earned less as high school graduates than measurably similar individuals who left after secondary school. No ability bias is found, as there is positive selection in both groups of the data.

Keane and Wolpin (1997) estimate a structural dynamic model of schooling, work and occupational choice and find that an extension of the human capital investment model does fit the data on school attendance, wages and work, this in turn, infers on future work decisions and wage patterns of young individuals. Particularly, when extending the human capital investment model allowing only for occupation-specific capital accumulation and unobserved endowments the model fails to explain persistence in occupational choice or the decline in schooling with age however, by including skills depreciation during unemployment, costs of job seeking and re-entry to school the model fits the data.

Britton et al. (2020) evaluate the impact of undergraduate education on lifetime earnings in the United Kingdom, this detailed study, among other results, suggests that the highest estimates of net discounted lifetime earnings major in Economics, Medicine, Mathematics and Law respectfully with figures around one million British pounds. The difference in income is amplified for individuals who attended selective universities over those who did not by about 0.4 million British pounds notably within top earners in men while average returns grow by eight-fold over the lifecycle.

Akerlof and Kranton (2002) review noneconomic papers and other approaches to the schooling literature. Since Schultz (1960) and Becker (1964) introduced the concept of human capital, economists have been apprehensive to debate resources devoted to education and their returns. John Bishop's (1998) questions the dichotomy of social norms that favour athletic over academic achievement, the latter, he argues, rewards only the individual, while the former awards the entire school. Bowles and Gintis (1976) argue that the U.S. educational system was designed to produce compliant workers and Kremer and Sarychev (2000) build a political economy model where schools inculcate their students with an ideology.

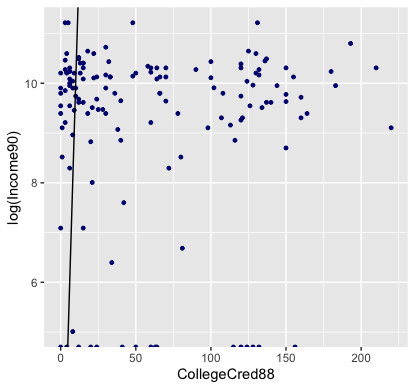
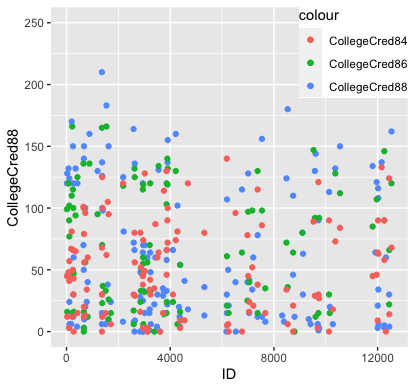
In the next section we introduce the NLSY79 data and discuss the data-cleaning and variable selection processes.

# 3. Data

## 3.1. Dataset

The National Longitudinal Survey of Youth Survey of 1979 or NLSY79 for short, is a longitudinal study of the lives of 12,686 Americans born between 1957 and 1964. When first interviewed, in 1979, their ages were between 14 and 22, two subsamples were dropped in the subsequent years and 9964 remained in the sample. The data is sparsely populated as the majority of participants do not answer all questions. The original data is collected each year from 1979 up to 2020 and variables are in the thousands mainly due to the repetition of certain questions concerning second, third up to fifth current job or college, typically only one current job is filled.

After examining these variables closely, it appears most necessary to select the ones with significant populations, in addition to filter for simultaneous availability of data for target variables. This drastically reduces the number of the selective cohort, as many who fill in one information do not fill in the other. To pursue the interest of the study, the variables we wish to select are income and its lags, college credit hours, occupational codes, a mental ability measure and a self-esteem score.



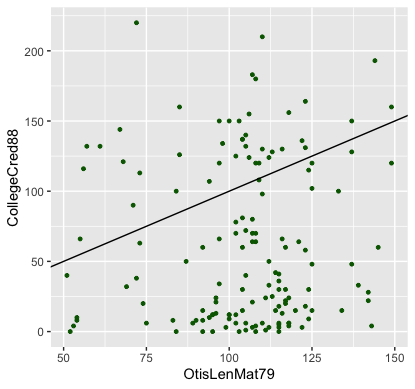
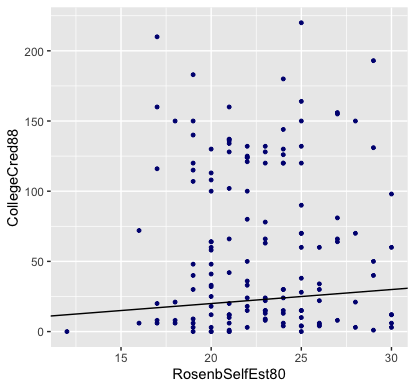
*Source: RStudio, Author’s work*

Figure 1: College credits accumulation by year (left), College on income(right)

We start with college credits where a subjective choice of the year 1988 ensures that all student would have passed 24 years of age, and thus are safely above the average age for graduation. In the US higher education system, a college year corresponds to 30 credits hours and a college degree to 120. The data is given initially in college credit hours.

From then, we focus on income after college, we interest ourselves with income in 1990 as a start and keep two lags, 1989 and 1988. As clarified by preliminary regressions, years prior to that are statistically insignificant. It is important to note that we are handling panel-data and that lags are not associated with the main variable beforehand. The income data is given in US dollars per year.

A mental ability test score is also collected, it is the variable with the least data points, nonetheless the Otis-Lennon test has the highest number of observations of all ability tests. This could be thought of as an “IQ” test taken in the initial interview of 1979. The test score is centred around 100 with values above 110-120 considered high and below 80-90 low.

*Source: RStudio, Author’s work*

Figure 2: College credits on mental ability (left) and self-esteem (right)

In addition, a Rosenberg self-esteem test is also taken the following year, the test is out of 30 and supposed to reflect the individual’s approval and satisfaction with themselves and their actions through a series of questions.

Occupation codes are collected for each year from 1979 to 1990. They come in the form of a three-digit code that correspond to a job denomination as devised in the 1970 CPS occupation code database. Each individual is also uniquely referred to by way of an identification number.

Data excluding occupation codes with no missing observation for income in 1990 and its two lags, college credits in 1988, the ability test and self-esteem scores is around 163 data points grouped by ID. Treatment of the data is entirely made in RStudio where this subgroup is referred to as ‘Clean’. In 1990, the cohort’s ages are between 26 and 33, which can be seen generally as a period of income stabilisation and “settlement” for most adults.

Figure 1 shows a stabilisation of college credits around 1988 contrasted by growing amounts during the acquisition phases in 1986 and 1984. A plot of log-income on college credits displays an almost inelastic relationship between the two, this initially suggests of an independent income variable of any educational background and that yearly log-income hovers typically around 10 with a unit standard deviation independently of college education.

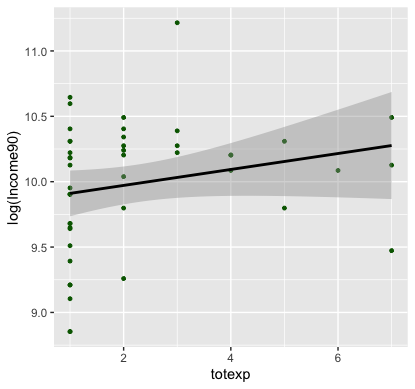
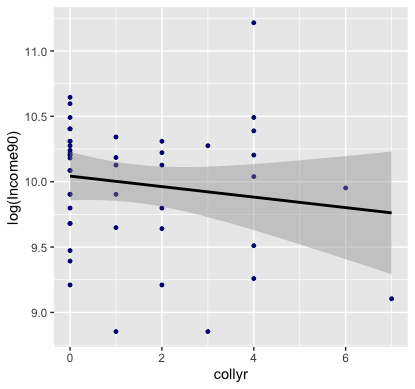
Figure 2, displays the relationship between college education and mental ability as reported by the Otis-Lennon test. A clear positive relationship exists which is evidence of self-selection occurring within the confounds of university. Mental ability is reflected in college credits by a 2:1 ratio corrupted by a high number of college abstinent. While self-esteem displays quasi-perfect elasticity with regards to education, it could be the lack of objective causality between self-esteem and college credits, although a slight positive linear relationship could suggest the same results in mental ability but on a smaller scale.

## 3.2. Derived data

Having gone through the process of data cleaning, we try to derive observations that are not explicitly stated in the dataset. Easier transformations of college credit hours, allow us to derive college education in years and college graduate status as a binary variable. We extract a high-income binary variable as well, defined as income higher than the national average at the time, high ability is also registered similarly.

Since no data is available on years of experience for the job worked in 1990, to extract this information we propose three functions. The first records the occupational code, the second is a binary variable that runs through each year and returns one if the code matches the 1990 digits, the third is the sum of the second for the years 1979 to 1989.

We have thus created a work experience variable expressed in years. This would be crucial as reliance on education alone will prove misleading. The extracted data is, by construction only available for data with all occupation codes from 1979 to 1990 filled up and thus only a subgroup of 46 agents has experience data. We refer to it as ‘Cleaner’ data.

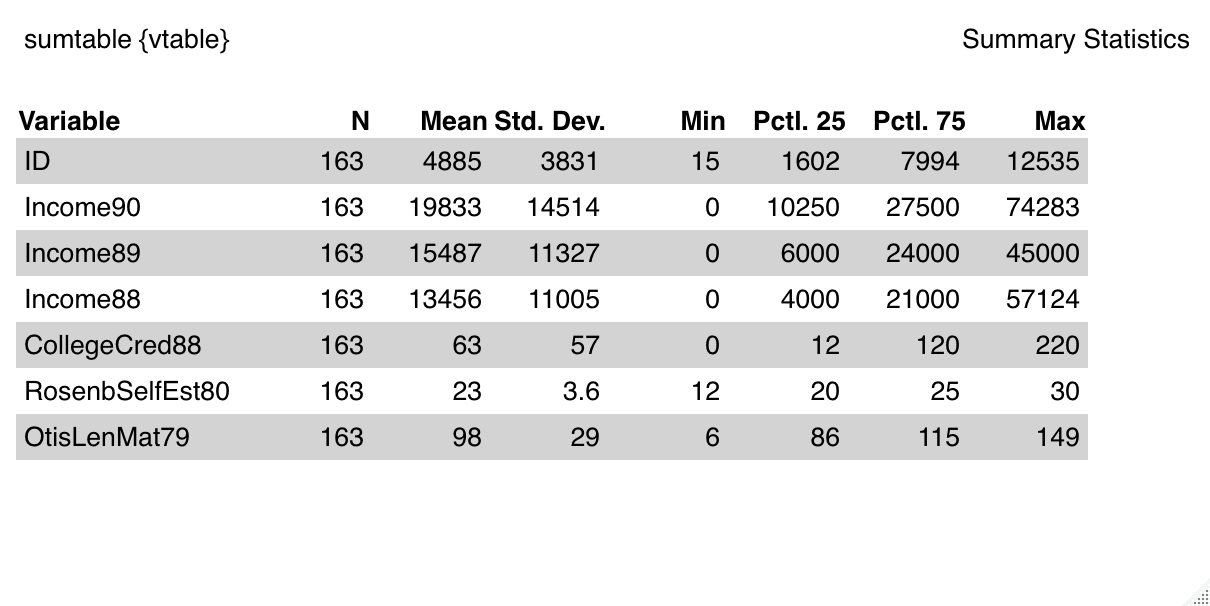
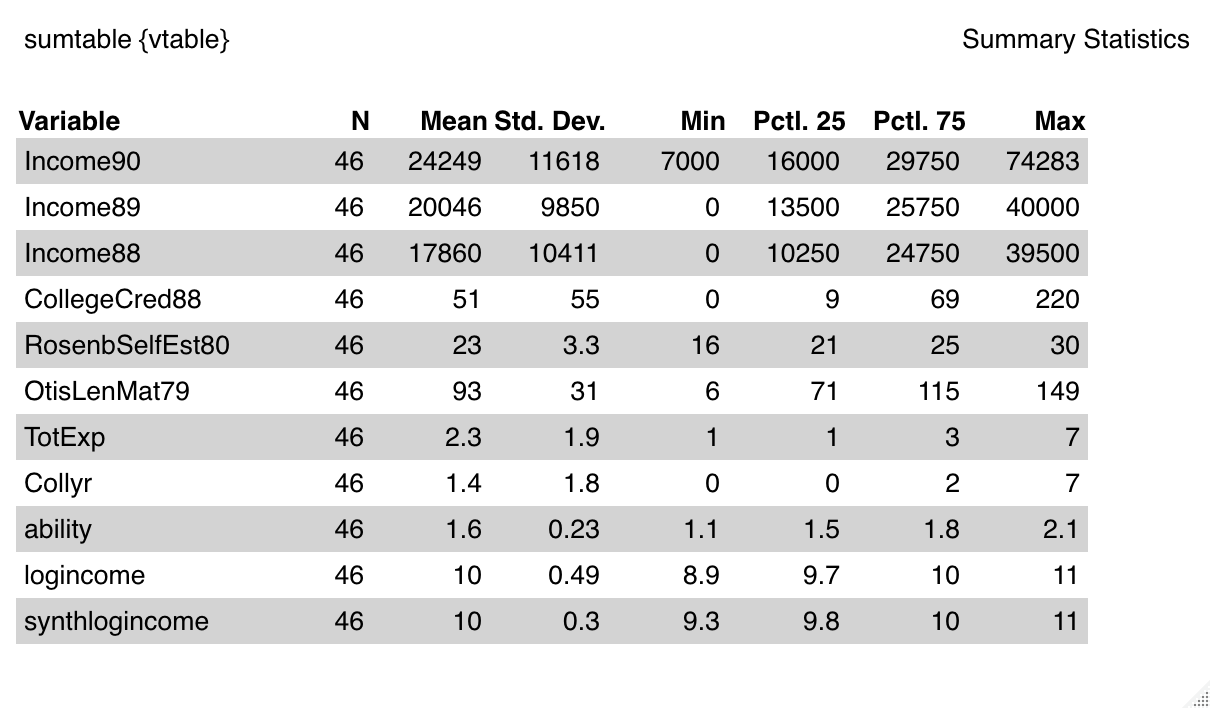
 

*Source: RStudio, Author’s work*

Figure 3: Income in logs on experience (left) and college (right) in year

Job-specific experience displays a positive relationship with log-income but disturbed by a significant number of novice workers without prior experience. While the college years discrete transformation is strangely enough negatively correlated with income, this result is in contradiction with that of college credits hours. This could be sample specific but could also be related to work experience especially in early stages of their careers, students with a high number of college years, more often than not have missed out on multiple years of work experience. Both experience and college discrete variables as displayed in Figure 3 will be included in the stochastic model along with mental ability and self-esteem.

## 3.3. Descriptive statistics

* *

*Source: RStudio, Author’s work*

Figure 4: Descriptive statistics table

The ‘Clean’ dataset with 163 observations excluding occupational variables shows a clear progression in average income each incremental year by roughly $2000 then $4800 respectively closely mimicked by the ‘Cleaner’ dataset with experience data although at lower volatility. Minimum income in 1990 in the large dataset is almost nil with in the occupational data is at $7000. An outlier in the income maxima of 1988 is not picked up in the smaller dataset.

The larger data contains on average individuals with higher mental ability by five points and higher college education by a margin of 12 credits while self-esteem scores are similar in both. Whether the presence of full information on occupational choice has anything to do with cohorts being both less able mentally and less educated is a matter of judgement but the experience hypothesis can apply in this context as workers with a full occupation history might have simply worked more and studied less.

On average, the occupation-inclusive data population has a little over two years of experience and close to one and a half years in college with a maximum of seven years and standard deviations of a little less than two years for each.

# 4. Model

## 4.1. Markov Decision Process

In mathematics, a Markov decision process or MDP for short, is a discrete-time stochastic control process, it is a modelling framework for decision making where outcomes are partly random and partly deterministic. MDPs are used extensively in optimization problems of dynamic programming. In reinforcement learning, problems concerned with how rational agents take actions in an environment so as to maximize cumulative reward, are often stated in the form of an MDP.

A Markov decision process is defined as a quadruple :

* State space
* Action space
* Reward matrix
* Probability transition matrix with:

A policy function is then defined as a mapping from the state space to the action space :

is defined as the optimal policy that maximises expected utility. It can be thought of as the path that leads to the highest reward given the states of the agent by taking a set of actions.

refers to the utility upon landing on state after taking action at state . It is given by the Bellman operator:

is the immediate reward and the second term is the discounted sum of future utilities reached from state by taking action . This is often referred to as policy evaluation.

The Markov property from which it derives its name is definitive of Markov chains:

MDPs are an extension of Markov chains, in the sense that if only one action exists for each state and all rewards are the same the model collapses to a Markov chain. Similarly, stochastic games generalise MDPs to more than one player, as well as dynamic situation where the environment itself changes due to agents’ choices (see. Shapley, 1953).

Since actions are completely determined by states through the mapping:

Solving for the optimal policy involves numerous methods, depending on the nature of the model, dynamic programming requires explicit values whereas Monte Carlo tree search and Reinforcement learning require a generative model. In the first, the output is the value function V, a discounted sum of rewards and policy that is a series of actions to follow.

4.2. Model specification

In our model, we chose the following settings:

The state space where corresponds to each state of the students’ college education categorised by the number of college credit hours:

* being no college education (i.e college credits less than 60)
* corresponds to an Associate’s degree
* corresponds to a Bachelor’s degree
* corresponds to a Master’s degree
* corresponds to further study PhD or post masters’

The action space where the agent simply decides to go into further study or stop and look for work.

* being to go into further study
* being to stop and work

A reward matrix is based on annual income generated from work with a certain level of education. It is defined as the average income of each subgroup categorised by college education is such that:

And the probability transition matrix is computed using annual statistics for the corresponding year, 1988, of college attendance in the US, it describes the proportion of American adults who have gone into each level of education. The transition probabilities satisfy the property:

This ensures, that no agent can go back to a previous state once arrived at his position -hence the Kronecker operator- and that Bayesian updating is introduced once the position changes - probabilities change whenever the agent moves closer to a position-. Also note that it is not allowed to jump two positions ahead at a time and only incremental change or stabilisation is allowed.

## 4.3. Reinforcement learning methods

### 4.3.1. Policy iteration

Policy iteration involves choosing an arbitrary policy for which we iteratively evaluate and improve the policy until convergence. We evaluate a policy by calculating the state value function V(s),

then the improved policy by using a one-step look-ahead to replace the initial policy .

Using this algorithm we produce a chain of policies, where each policy is an improvement over the previous one. We conduct policy evaluation and policy improvement steps until the policy doesn’t improve anymore.

### 4.3.2. Value iteration

In value iteration, we do not use the policy but rather the value function that is evaluated and improved upon until convergence.

It starts with an initial guess , often zero, and an iteration of evaluation and improvement is repeated until converges.

### 4.3.3. Q-learning

In reinforcement learning, Q-learning is a model-free algorithm that learns the expected rewards for an action taken in a given state. For any finite MDPs, there exists an optimal policy to be derived. Q refers to the quality of the state-action combination with respect to maximizing future rewards. At its core it is a value iteration updating system that uses a weighted average of the present value and the upcoming value.

The weight is the learning rate, it determines the step size at each iteration and represents to which extent the agent uses exclusively prior information, , or current data only, , in practice it is commonly set as 0.1. The algorithm ends with a terminal state where Q converges.

The most important distinction to make is that reinforcement learning can solve Markov-Decision processes without explicit specification of the transition probabilities. Unlike in value and policy iteration, the transition probabilities are achieved by a repeated simulation from a uniformly random initial state.

## 4.4. Stochastic model of rewards

In later sections, we would like to know how the process differs with the absence of known rewards, i.e., we substitute the deterministic reward matrix for a stochastic model of log-income based on college education, work experience and a couple of ability measures. For that we construct a log-linear model as follows:

For this model, we make us of the experience discrete variable derived earlier, a transformation of college credits into years of education and an ability measure.

Since college credits do not yield any significant statistical result, we transform the data into years of schooling with increments of one year for each 30 credits. While the ability measure is reflective of the Self-esteem and Mental Ability scores, we initially include them in the regression model to get relative weights with regards to the dependant variable and each other as the weighted sum would represent ability that matches one-for-one log income. As they do not vary over time, including the fitted values of the ability components back as one variable instead of two, using weights from the initial regression improves the fit by two percent.

# 5. Findings

## 5.1. Complete information case

We start with a Markov Decision process with complete information, where the agent forms an expectation of his reward given income data of the cohort sorted by educational level.

He is also aware of the state and action spaces, i.e., he knows at which state he is and what possible actions he can take.

The transition matrix ‘Pp’ is also provided as estimates of national college attendance are used and updated via Bayes law to include previous levels of education. It is made up of five 5x5 matrices one for each state. The reward matrix ‘Rr’ is 5x5. They both reduce to one 5x5 and 5x1 respectively for a given policy.

We evaluate a policy of ‘pursuing further study’, this is simply choosing to go into further education each time the agent reaches a new level. At the fifth and final state, the agent is in a static state with virtually no other action than to go out of study.

|  |
| --- |
| mdp\_computePpolicyPRpolicy (Pp, Rr, policy = c (1,2,3,4,5))  [[1]]  [,1] [,2] [,3] [,4] [,5]  [1,] 0.6578947 0.3421053 0.0000000 0.0000000 0.0000000  [2,] 0.0000000 0.6341463 0.3658537 0.0000000 0.0000000  [3,] 0.0000000 0.0000000 0.6818182 0.3181818 0.0000000  [4,] 0.0000000 0.0000000 0.0000000 0.7777778 0.2222222  [5,] 0.0000000 0.0000000 0.0000000 0.0000000 1.0000000  [[2]]  [1] 20214.42 15962.53 22053.19 21567.57 19500.00 |

*Source: RStudio, Author’s work*

Figure 5: Policy evaluation of ‘pursuing further study’

It is important to note that the choice of classification for level four, postgraduate study, influences how returns are displayed, as one agent with 193 college credits earns an income high enough to distort the sample whether it be that of doctors consisting of three agents only or that of postgraduates. We chose to keep this agent and by extension others below his college credits amount at postgraduate study level simply because they do not fulfil total credit requirements for a doctor’s classification.

Above is the transition and reward matrices for the selected policy. One can clearly see that reward is maximised for this particular sample at graduate level with a subtle drop at postgraduate level and doctorate level. A more interesting result is that the lowest earning potential is that of Associates’ degree holders rather than those who did not attend college.

A first explanation would be that college graduates have managed to acquire more years of experience compared to postgraduates which might have increased their earnings, the same could be said for those who skipped college altogether with respect to those who earned an associates’ degree.

Another explanation is that Associates’ holders send perhaps the wrong signal to employers that they are unable to continue college education until full achievement and thus employers simply do not want to take such a risk. As a result, these students not only are unable to get high income jobs but are also liable to the burden of its costs.

|  |
| --- |
| mdp\_bellman\_operator (Pp, Rr, 0.95, Vprev = c(0,0,0,0,0))  $V  [1] 20214.42 22053.19 22053.19 21567.57 19500.00  $policy  [1] 1 3 3 4 5 |

*Source: RStudio, Author’s work*

Figure 6: Optimal policy and Bellman operator of ‘pursuing further study’

By computing the Bellman operator for any discount factor, we arrive at the optimal policy to adopt at each state. Most intuitively, if the agent is at associates level/state he must take action to jump to graduate state. And should stay in his position if at other states. This result is the same independently of the discount factor. We seek to question this result by iterating this policy until convergence.

|  |
| --- |
| mdp\_value\_iteration (Pp, Rr, 0.95, epsilon=0.1, max\_iter = 40, V0=c(0,0,0,0,0))  $V  [1] 379478.4 384381.8 376048.2 347798.6 339880.3  $policy  [1] 1 3 4 4 5  $iter  [1] 40 |

*Source: RStudio, Author’s work*

Figure 7: Policy iteration for patient agents

The results are indeed different, using the value iteration method the MDP optimal policy yields somewhat more conventional results. If the agent has no college education the iteration advises to stay in that state. If one holds an associate’s then he should pursue a full college degree. Subsequently, if one holds a college degree the algorithm suggests enrolling in postgraduate study. And if one holds a postgraduate degree then it is best if he stays at postgraduate level, obviously if one holds a doctorate no action in that regard is to be taken.

Keep in mind that the initial algorithm took 257 iterations to converge and yields the same results regarding optimal policy as for 40 iterations. As income data is yearly, we deliberately keep 40 iterations as a proxy for the active life from early adulthood to retirement this also give us a discounted cash flow amount with which to assess lifetime earnings.

Another important point in how much the agent in question values current earning as opposed to future income, the results above are valid for agent with a discount factor strictly above 0.7 up to, and excluding, 1 to sustain convergence of the series.

|  |
| --- |
| mdp\_value\_iteration (Pp, Rr, discount=0.69, epsilon=0.1, max\_iter = 40, V0=c(0,0,0,0,0))  $V  [1] 67771.22 71138.61 69574.14 67364.97 62902.59  $policy  [1] 1 3 3 4 5  $iter  [1] 31 |

*Source: RStudio, Author’s work*

Figure 8: Policy iteration of for less patient agents

For less patient agents results are indeed different, specifically for graduate degree holders, which based on this characteristic, would stay at graduate level rather that pursue further study. The value function is reduced five-fold to reflect the higher discount. This policy applies to all agents with discount factors strictly below 0.7 up to and excluding the limit of zero for which it collapses to a single Bellman operator computation with some change and the number of iterations necessary for convergence reduces up to two accordingly.

## 5.2. Incomplete information case

### 5.2.1. Hidden transition probabilities

If we are to omit the transition matrix, then the agent ought to implement some simulation method to reproduce transition probabilities. The most appropriate approach is to use Q-learning, as mentioned before this method does not need an explicit transition probability matrix, but rather an initial guess to start with. The algorithm entries are the initial probability guess which we set at the identity matrix, the reward matrix, a discount factor and the iterations start and are fixed at 10000.

We compile a set of policies depending on the agent’s type from low to high valuation of future returns with 11 discount factors from 0.05 to 0.95. In parallel, we reproduce the same procedure but introduce the full probability transition matrix so that comparison can be made with the complete information case. The initial and final states are unaltered, agents no matter their type will stay put if at the initial stage or at the final.

|  |
| --- |
| Qfunct<-function(a){mdp\_Q\_learning(PIdt, Rr, a)$policy}  sapply(odisc, Qfunct)  # With P identity matrix  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 1 1 1 1 1 1 1 1  [2,] 3 2 3 2 3 2 2 3 2 3 3  [3,] 3 4 4 3 4 3 4 3 3 3 3  [4,] 4 4 4 5 4 5 5 4 4 4 4  [5,] 5 5 5 5 5 5 5 5 5 5 5  # With P given as Pp  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 1 1 1 1 2 1 1 1  [2,] 2 2 2 3 3 2 3 3 3 3 3  [3,] 3 4 4 4 4 4 3 3 3 3 3  [4,] 4 5 4 4 5 4 5 5 5 4 5  [5,] 5 5 5 5 5 5 5 5 5 5 5 |

*Source: RStudio, Author’s work*

Figure 9: Q-learning optimal policy by discount factors

The more interesting results are in the middle: At associate’s level, it seems as if the discount factor alone does not give a clear indication about the optimal action, except for high enough levels of patience (0.9 and 0.95) where the agent should pursue a full college degree. This is contrasted by the full information case where almost all types above 0.3 would optimally jump to the next state.

For college graduates, the simulation suggests for patient agents (0.7 and above) to stay in their state while for less patient ones, mixed signals are weakly dominated by a further study action. This is partly due to the value function dropping slightly at postgraduate level and thus suggesting comparatively unfavourable outcomes. These same results are more distinctly displayed in the full information case.

Postgraduates’ optimal decision is to stay at their level, to the exception of three intermediate ‘types’. This again is due to the subsequent small drop in the corresponding value function. This contrasts with the complete information case that suggests, in most part, pursuing further study.

Rewards at both ends of the educational spectrum, by nature of the underlying data are either higher than in the subsequent state -as is the case for non-college educated- or incrementally lower than the previous state but with a significant transition probability and no exit strategy that they constitute equilibrium states for the agent.

A potential equilibrium for patient agents specifically is that of a full college degree confirmed by both cases. The postgraduate degree state might be clearly in equilibrium in the simulation but suggests the opposite with predetermined transition probabilities.

### 5.2.2. Hidden rewards

In this section, the agent does not know about the reward each state procures, he/she therefore needs to make use of other data to model income. The regression results of a log-linear stochastic model for rewards incorporates college education, work experience and the ability measure. All elements are statistically significant at the 5% level, with college years slightly at the outer edge with the rest also statistically significant at the 1% level, except the square of education years sitting outside of the 1% interval by a small margin. The model has an adjusted fit of 0.3 and is tested for various anomalies (heteroskedasticity, multicollinearity, non-linearity...) using standard tests.

|  |
| --- |
| lm(formula = log(Income90) ~ Collyr + TotExp + I(TotExp^2) + Ability, data = Occtest)  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 8.00230 0.46124 17.350 < 2e-16 \*\*\*  Collyr -0.06945 0.03488 -1.991 0.053140 .  TotExp 0.39835 0.14608 2.727 0.009366 \*\*  I(TotExp^2) -0.05059 0.01957 -2.585 0.013398 \*  Ability 1.00004 0.27370 3.654 0.000727 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.4097 on 41 degrees of freedom  Multiple R-squared: 0.3647, Adjusted R-squared: 0.3028  F-statistic: 5.885 on 4 and 41 DF, p-value: 0.0007741 |

*Source: RStudio, Author’s work*

Figure 10: Log-linear model of income on college, experience and ability

College education, in years, seems to decrease log-income by a factor of 0.06. While each year of work experience increases log-income by 0.39, while at the quadratic level decreases it by a factor of 0.05. A one-point increase in the self-esteem and mental ability scores results in 0.05 and 0.003 positive change in log-income. To improve the fit, and reduce the regressors we are able to record them both with the corresponding weights as an ability measure. (See appendix)

We record the fitted values of this model and proceed to the MDP process: first we stratify by college education the rewards as the synthetic log-income and include the original log-income for reference. The stratification reproduces very close average log-income for each educational level and conserves most comparative results in the reward matrix.

|  |
| --- |
| #Synthetic  SLRr  [,1] [,2] [,3] [,4] [,5]  10.01856 9.92047 10.04685 9.616076 9.511965  #Original  LRr  [,1] [,2] [,3] [,4] [,5]  10.02611 9.804433 10.15828 9.952278 9.10498 |

*Source: RStudio, Author’s work*

Figure 11: Synthetic and original log-income rewards

We then apply a policy iteration method to solve for the epsilon-optimal policy. It is found after 92 iterations but yields the same policy results as for 40 iterations, which we keep instead.

|  |
| --- |
| #Synthetic  mdp\_value\_iteration (Pp, SLRr, 0.95, epsilon=0.1, max\_iter = 40, V0=c(0,0,0,0,0))  $V  [1] 175.0388 175.1142 167.8117 166.1900 165.7912  $policy  [1] 1 3 4 4 5  #Original  mdp\_value\_iteration (Pp, LRr, 0.95, epsilon=0.1, max\_iter = 40, V0=c(0,0,0,0,0))  $V  [1] 176.7039 177.0564 173.5215 161.9425 158.6976  $policy  [1] 1 3 4 4 5 |

*Source: RStudio, Author’s work*

Figure 12: Value iteration for synthetic and original log-income rewards

The synthetic data suggests the same policy as that of the original income data. The agent without prior knowledge of his income can form an expectation about his reward by resorting to his educational attainment, work experience and ability data. This in turn would yield the Markov decision process computations in the same fashion as in the complete information case. The discount factor parametrisation also returns the same results as in the original case.

### 5.2.3. Cost evaluation

We would like to introduce a cost function related to college tuition fees as the unitary price of a year of schooling for the corresponding period by the number of years. This would enable us to clearly identify the optimal choice given realistic circumstances. In 1988, the tuition for a private non-profit four-year institution is $15,160 per annum, while a public four-year institution costs $3,190 p.a. These costs have increased by a markup of 163% in 30 years (Martin, 2017).

The distribution is such that 77.4% of all postsecondary students attend public institutions and 22.6% of postsecondary students attend private institutions, 80.0% of which are non-profit schools (Hanson, 2022). We evaluate the costs first using a weighted average of $5895.22 p.a, then treat each case separately by stacking costs against future discounted income using the Bellman operator.

|  |
| --- |
| #Public education  Bellman - cf[1,]  [,1] [,2] [,3] [,4] [,5]  [1] 22053.192 15673.192 9293.192 6103.192 -3466.808  Bellman - cff[1,]  [1] 22053.192 15673.192 9293.192 6103.192 6103.192  #Private non-profit education  Bellman - cf[1,]  [1] 22053.192 -8266.808 -38586.808 -53746.808 -99226.808  Bellman - cff[1,]  [1] 22053.192 -8266.808 -38586.808 -53746.808 -53746.808 |

*Source: RStudio, Author’s work*

Figure 13: Cost function

For the weighted average costs one can see positive benefits for non-school attendants and associate’s holders with a slight loss at the Bachelor’s level. If the cost were public education specific however, students are able to retrieve positive benefits from education at all levels including Bachelor’s, Master’s and fully-funded Doctorates. A self-funded PhD is the only decision that is cost-benefit negative. On the complete opposite of the spectrum, private education results suggest no economic benefit even at Associates’ level and that a rational agent faced with this information would simply abstain from higher education.

The results shown for public education and fully-funded doctorates are the clearest indication of the importance of educational attainment in lifetime earnings potential. While the private education results, highlight the significant entry barriers that can potentially prevent able students from self-identifying in a higher education program.

### 5.2.4. Synthetic model

At this point we are able to combine the methods used, in cases where more than one input is omitted to the agent. We can evaluate these results against costs in the same manner.

|  |
| --- |
| # Policy iteration net value  # 0.05 0.1 0.3 0.5 0.7 0.9 0.95  [,1] [,2] [,4] [,6] [,8] [,10] [,11]  [1,] 23632.132 24958.573 32173.531 45206.54 75749.891 229251.81 459943.7  [2,] -6022.062 -4672.177 2655.773 15846.08 46623.470 200510.41 431340.8  [3,] -36482.428 -35300.052 -29071.309 -18504.33 4109.505 102980.56 241791.1  [4,] -60024.301 -59168.536 -54549.911 -46331.34 -27475.021 64350.45 200292.3  [5,] -107047.858 -106257.183 -101964.950 -94238.93 -76211.550 13925.35 149130.7 |

*Source: RStudio, Author’s work*

Figure 14: Synthetic model value function net of costs using policy iteration in private education

Above is an application of policy iteration to the stochastic model evaluated at cost as a solution to a case of omitted rewards with frictional costs where transition probabilities are known.

The second model below, is the result of applying the Q-learning method to the stochastic model in order to deal with absent transition probability and unknown rewards matrices. The results replicate the original case with full information.

|  |
| --- |
| # Q-learning net value  # 0.05 0.1 0.3 0.5 0.7 0.9 0.95  [,1] [,2] [,4] [,6] [,8] [,10] [,11]  [1,] 23620.144 24932.374 32055.908 44878.25 74704.59 220322.587 352289.0  [2,] -8906.783 -4672.177 2655.773 10365.11 46623.35 192979.372 385751.0  [3,] -36342.062 -34992.177 -27664.227 -19091.28 16262.23 165576.192 169620.2  [4,] -61567.858 -59128.817 -54365.621 -48758.93 -30731.59 68482.458 167561.9  [5,] -107047.858 -106257.191 -101964.950 -94238.93 -76219.93 -1813.124 48886.1 |

*Source: RStudio, Author’s work*

Figure 15: Synthetic model value function net of cost using Q-learning in private education

This model only requires the agent’s type, knowledge of his state and the actions he can take to paint a full picture of the potential rewards, simulate the probabilities of moving to each state and evaluate against entry costs. It is implicit to revert from logarithms in synthetic income using exponentiation to obtain valid results.

|  |
| --- |
| # Policy iteration  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 1 1 1 1 1 1 1 1  [2,] 3 3 3 3 3 3 3 3 3 3 3  [3,] 3 3 3 3 3 3 3 3 3 3 3  [4,] 4 4 4 4 4 4 4 4 4 4 4  [5,] 5 5 5 5 5 5 5 5 5 5 5 |

|  |
| --- |
| # Q-learning  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 2 2 1 1 2 1 1 2  [2,] 2 2 2 2 2 2 3 3 3 3 3  [3,] 3 3 3 3 3 4 3 4 3 3 4  [4,] 5 4 5 5 5 4 4 5 5 4 5  [5,] 5 5 5 5 5 5 5 5 5 5 5 |

*Source: RStudio, Author’s work*

Figure 16: Optimal policy by type in the synthetic model

### 5.2.5. Evidence of reversal in self-identification

Since the final model output contains multiple states with negative rewards, we would like to explore a scenario where agents are allowed to reverse at negative value. This imitates a situation where the agent is devising a strategy that responds to negative output by reverting to precedent states. We relax the assumption in the Kronecker operator used up to this point and allow for movement backwards if and only if values are negative.

This move would be equivalent in reality to a profit-maximising agent self-identifying as less educated simply by stating a lower academic achievement as his highest in order to earn more. This is a case where the incentive compatibility constraint breaks down and the agent no longer has a motive to behave in a manner consistent with the optimal solution.

With private costs, only very patient individuals (0.90.95) would gain from enrolling in postgraduate study, patient agents (0.70.8) would identify as bachelor’s holders to maximise income, less patient ones (0.30.6) at associates and impatient agents (0.050.2) would not enrol. In these cases, agents would self-identify with lower credential if at later stages of the educational process. This is a counter-intuitive result that can be related to the stochastic model with negative effects of college years on income or simply evidence of an educational system where high-types no longer have an incentive to be recognised as such.

Public costs however still allow for greater participation with impatient individuals pursuing up to bachelor’s degrees, less patient opting for master’s and patient individuals for doctorates especially when fully-funded.

|  |
| --- |
| # Q-learning Modified optimal policy with reversion  # Private costs  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 2 2 1 1 2 1 1 2  [2,] 1 1 1 2 2 2 3 3 3 3 3  [3,] 1 1 1 2 2 2 2 3 3 3 4  [4,] 1 1 1 2 2 2 2 3 3 4 5  [5,] 1 1 1 2 2 2 2 3 3 4 5 |

|  |
| --- |
| #Public costs  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 2 2 1 1 2 1 1 2  [2,] 2 2 2 2 2 2 3 3 3 3 3  [3,] 3 3 3 3 3 4 3 4 3 3 4  [4,] 3 3 4 4 4 4 4 5 5 4 5  [5,] 3 3 4 4 4 5 5 5 5 5 5 |

|  |
| --- |
| # Public costs funded PhD  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 1 1 2 2 1 1 2 1  [2,] 2 2 2 3 3 2 2 3 2 3 3  [3,] 3 3 3 4 4 3 4 4 4 3 4  [4,] 5 4 4 5 5 5 4 5 4 5 5  [5,] 5 5 5 5 5 5 5 5 5 5 5 |

*Source: RStudio, Author’s work*

Figure 17: Q-learning modified optimal policy with reversion by type with private, public self-funded and fully-funded PhD

# 6. Synthesis

In this study we have been able to decorticate the investment decision of going into university for eligible prospective students. Questions on the optimal duration of study, when to stop and not to stop have been identified and answered.

The study of education as an investment decision through the lenses of a Markov decision process clarifies the somewhat obscure measure of education and its impact on lifetime earnings. Using the baseline case, with full information we are able to lay out a reference for our policy and its rewards. One can clearly retain that a policy equilibrium does not exists at the Associate’s level and that it is an unstable point. Also, that policy iteration suggests an equilibrium at the Bachelor’s level for less patient individuals and at the Master’s level for patient agents.

The second part of the study involves omitting key inputs in the Markov decision process to account for more realistic circumstances a student might find himself in. The study with incomplete information also displays additional methods of extracting and analysing the data. First, we omit the transition matrix, the agent is aware of his state and the action available to him but knows not the probability of attaining a certain educational level. To solve this, we implement a method of reinforcement learning, Q-learning, to simulate the data without the required input of transition probabilities. The parametrisation by increments of the discount factor highlights key differences with respect to the complete information case especially at postgraduate level where the latter suggests pursuing further study and the simulation to stay at the same level. Unsurprisingly the full information case has more clearcut policy suggestions compared to the simulation.

Second, we omit the reward matrix, the agent knows not what rewards each state procures and thus in order to make an informed decision we require a stochastic model for log-income based on educational attainment, work experience and ability, mental and self-esteem. Recoding the fitted values produces a synthetic reward matrix to which we subject the same analysis. In the absence of explicit rewards, the agent forms an expectation using observables he can collect and as a result can reproduce the same policy path as in the initial case.

Third, we relax the assumption of frictionless transactions and introduce a cost function to reflect the tuition fees involved in higher education which we set against future income embodied in the Bellman operator. A distinction is made between public and private non-profit institutions and the underlying cost per annum. For public universities, the economic benefits outweigh the cost at all levels when Doctorates are fully funded and up to the Master’s level when self-funding. In contrast, private education results in a loss.

# 7. Conclusion

Throughout the paper, we have introduced a novel approach to model prospective student’s decision to join higher education and the optimal quantity to undertake. We departed from the baseline model to include cases of omitted information about the agent’s reward and its transition probabilities and evaluated against tuition costs involved. In doing so, we not only display the robustness of the framework as a novelty to the field but more importantly its usefulness to evaluate a lifetime investment decision in higher education and point towards the right policy to undertake. This also highlights the relevance of reinforcement learning methods in producing a great deal of insight in economic evaluation and discrete choice modelling.

# References

* Akerlof, George A, and Rachel E Kranton. “Identity and Schooling: Some Lessons for the Economics of Education.” *Journal of Economic Literature*, vol. 40, no. 4, Dec. 2002, pp. 1167–1201, https://doi.org/10.1257/.40.4.1167. Accessed 18 Aug. 2023.
* Albrecht, James W. “A Procedure for Testing the Signalling Hypothesis.” *Journal of Public Economics*, vol. 15, no. 1, Feb. 1981, pp. 123–132, https://doi.org/10.1016/0047-2727(81)90057-8. Accessed 18 Aug. 2023.
* Arrow, Kenneth J. “Higher Education as a Filter.” *Journal of Public Economics*, vol. 2, no. 3, July 1973, pp. 193–216, https://doi.org/10.1016/0047-2727(73)90013-3. Accessed 18 Aug. 2023.
* Autor, David H. “Work of the Past, Work of the Future.” *American Economic Association*, 2019, www.aeaweb.org/articles?id=10.1257/pandp.20191110.
* Bade, Sophie. “Nash Equilibrium in Games with Incomplete Preferences.” *Economic Theory*, vol. 26, no. 2, Aug. 2005, pp. 309–332, https://doi.org/10.1007/s00199-004-0541-1. Accessed 18 Aug. 2023.
* Bellman, Richard. “A Markovian Decision Process.” *Journal of Mathematics and Mechanics*, vol. 6, no. 5, 1957, pp. 679–684, www.jstor.org/stable/24900506.
* Bertsekas, Dimitri. *Dynamic Programming and Optimal Control 3rd Edition, Volume II Chapter 6 Approximate Dynamic Programming*. 2011.
* Blundell, Richard, et al. “Evaluating the Effect of Education on Earnings: Models, Methods and Results from the National Child Development Survey.” *Journal of the Royal Statistical Society Series A: Statistics in Society*, vol. 168, no. 3, 10 Mar. 2005, pp. 473–512, https://doi.org/10.1111/j.1467-985x.2004.00360.x. Accessed 18 Aug. 2023.
* Britton, Jack, et al. *The Impact of Undergraduate Degrees on Lifetime Earnings*. 2020.
* Bureau, US Census. “Current Population Survey Tables for Personal Income.” *The United States Census Bureau*, 2023, www.census.gov/data/tables/time-series/demo/income-poverty/cps-pinc.html.
* Camerer, Colin, and Teck-Hua Ho. *Behavioral Game Theory Experiments and Modeling*. 2014.
* Clark, Kenneth, and Martin Sefton. “Repetition and Signalling: Experimental Evidence from Games with Efficient Equilibria.” *Economics Letters*, vol. 70, no. 3, Mar. 2001, pp. 357–362, https://doi.org/10.1016/s0165-1765(00)00381-5. Accessed 18 Aug. 2023.
* Feinberg, Eugene A., and Shwartz Adam. “Handbook of Markov Decision Processes: Methods and Applications.” *Springer*, 2002.
* Fianu, Sefakor, and Lauren B. Davis. “A Markov Decision Process Model for Equitable Distribution of Supplies under Uncertainty.” *European Journal of Operational Research*, vol. 264, no. 3, Feb. 2018, pp. 1101–1115, https://doi.org/10.1016/j.ejor.2017.07.017. Accessed 18 Aug. 2023.
* Fourny, Ghislain. “Perfect Prediction in Minkowski Spacetime: Perfectly Transparent Equilibrium for Dynamic Games with Imperfect Information Working Paper.” *ETH Library* , https://doi.org/10.3929/ethz-b-000341683. Accessed 18 Aug. 2023.
* Hanson, Melanie. “College Enrollment & Student Demographic Statistics.” *Education Data Initiative*, 2022, educationdata.org/college-enrollment-statistics#. Accessed 18 Aug. 2023.
* Jonckheere, A. R., et al. “Stochastic Models for Learning.” *Biometrika*, vol. 43, no. 1/2, June 1956, p. 237, https://doi.org/10.2307/2333607. Accessed 18 Aug. 2023.
* Kallenberg, M. “Finite State and Action MDPS.” *International Series in Management Science/Operations Research*, 1 Jan. 2003, pp. 21–87, https://doi.org/10.1007/978-1-4615-0805-2\_2. Accessed 18 Aug. 2023.
* Keane, Michael P., and Kenneth L. Wolpin. “The Carrer Decisions of Young Men.” *Journal of Political Economy*, 1997, www.jstor.org/stable/10.1086/262080.
* Martin, Emmie. “Here’s How Much More Expensive It Is for You to Go to College than It Was for Your Parents.” *CNBC*, CNBC, 29 Nov. 2017, www.cnbc.com/2017/11/29/how-much-college-tuition-has-increased-from-1988-to-2018.html.
* Munos, Remi. *Efficient Resources Allocation for Markov Decision Processes*. 2001.
* Ng, Ritchie. “Deep Learning Materials by Deep Learning Wizard.” *GitHub*, 16 Aug. 2023, github.com/ritchieng/deep-learning-wizard/. Accessed 18 Aug. 2023.
* Philippon, Thomas, and Jan Eeckhout. *Book Review Journal of Economic Literature \* the Great Reversal*. 2020.
* Pritchett, L. “Where Has All the Education Gone?” *The World Bank Economic Review*, vol. 15, no. 3, 1 Oct. 2001, pp. 367–391, https://doi.org/10.1093/wber/15.3.367. Accessed 18 Aug. 2023.
* Puterman, Martin L., and Moon Chirl Shin. “Modified Policy Iteration Algorithms for Discounted Markov Decision Problems.” *Management Science*, vol. 24, no. 11, July 1978, pp. 1127–1137, https://doi.org/10.1287/mnsc.24.11.1127. Accessed 23 Oct. 2019.
* Riley, John. “Competitive SignaIling.” *Journal of Economic Theory*, vol. 10, 1975, pp. 174–186. Accessed 18 Aug. 2023.
* Rothschild, Michael, and Joseph Stiglitz. “Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information.” *The Quarterly Journal of Economics*, vol. 90, no. 4, Nov. 1976, p. 629, https://doi.org/10.2307/1885326. Accessed 18 Aug. 2023.
* Shapley, L. S. “Stochastic Games.” *Proceedings of the National Academy of Sciences*, vol. 39, no. 10, 1 Oct. 1953, pp. 1095–1100, https://doi.org/10.1073/pnas.39.10.1953. Accessed 18 Aug. 2023.
* Spence, Michael. “Job Market Signalling.” *The Quarterly Journal of Economics*, vol. 87, no. 3, 1973, pp. 355–374. Accessed 18 Aug. 2023.
* Stiglitz, Joseph. “The Theory of “Screening,” Education, and the Distribution of Income.” *Source: The American Economic Review*, vol. 65, no. 3, 1975, pp. 283–300. Accessed 18 Aug. 2023.
* Sutton, Richard S., and Andrew G. Barto. “Reinforcement Learning: An Introduction.” *MIT Press*, 1998, www.andrew.cmu.edu/course/10-703/textbook/BartoSutton.pdf.
* Walker, Ian , and Yu Zhu. *The Impact of University Degrees on the Lifecycle of Earnings: Some Further Analysis*. Department for Business Innovation and Skills, 2013.
* Willis, Robert, and Sherwin Rosen. “Part 2: Education and Income Distribution.” *Source: Journal of Political Economy*, vol. 87, no. 5, 1979, pp. 7–36. Accessed 18 Aug. 2023.

# Appendix

**A. Other models of income, education and ability**

In order to compute the MDP, we allow for different approaches to model income, education and ability. We attempt linear and logit models and would keep if needed a deterministic model based on average only. This latter case would allow for the decision-maker the choice of policy given prior knowledge of future income. In the other models, the agent need not know his future income but only determinants of his income to model his reward.

4.3.1. Log-linear model with college education as a dependant variable

Where is log income, we regress income on its two lags, college credits, ability and self-esteem. Only income and its two lags are statistically significant for the NLSY79 cohort and thus income is an AR(2).

For college credits, we regress college credits on mental ability and self-esteem scores to allow for self-selection of high types into high credit hours achievement as evidenced by the signalling model of education. Surprisingly enough, the model has no statistically significant elements apart from its constant, which might suggest recalling a deterministic model instead.

4.3.2. Log-linear model with income lags

Where is log-income, the model is statistically significant for both ability and esteem and is evidence of high- type individuals ending up earning higher income. The model however doesn’t include college credit hours -or log-credits- as it is statistically insignificant. The fit is very poor 0.07 adj-R. The model with income lags, however has high-multicollinearity and the context in which the study occurs, especially in the case hidden rewards doesn’t allow prior knowledge of lag income. For both these reasons would make sense to discard lag income for relevance.

4.3.3. Logit model for high income and college graduation

In this model the probability of earning above a certain high-income threshold is regressed on log college credits. Another approach is to model the probability of being a college graduate on ability and self-esteem. Both these logit models are statistically insignificant with a fit close to zero.

**B. Q-learning output**

|  |
| --- |
| sapply(odisc, VQfunct)  # With P identity matrix  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 21278.33 22460.46 25268.02 28877.74 33690.70 40428.55 50536.04 67092.08 100934.66 196369.4 322965.4  [2,] 16802.67 17736.15 19953.17 31504.56 36755.02 44106.38 55075.02 73509.12 077855.99 156191.0 269134.7  [3,] 23213.89 24503.55 26959.46 31504.56 35945.47 43135.14 55132.87 73187.10 107811.53 199950.4 306874.0  [4,] 22702.71 23963.97 26959.46 27857.14 32500.00 43135.14 53918.93 64999.96 097493.52 204991.4 314109.4  [5,] 20526.32 21666.67 24375.00 27857.14 32500.00 39000.00 48749.98 64999.99 097342.91 191892.9 349702.5  # With P given as Pp  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 16802.66 9541.70 22255.04 13527.54 18560.47 19535.69 10568.47 27668.12 18888.63 32703.71 19283.50  [2,] 16243.49 24503.55 27566.49 14500.40 21268.84 44106.38 29700.81 73510.63 33604.49 73846.94 56794.75  [3,] 22242.06 22304.28 25102.42 30810.82 35945.95 43135.14 53918.93 55817.70 74739.20 151920.03 156258.16  [4,] 22503.25 23713.35 26411.82 27857.14 32500.00 32827.97 48750.00 65000.00 94790.98 194989.39 299909.64  [5,] 20526.32 21666.67 24375.00 27857.14 32500.00 39000.00 48750.00 64999.80 97500.00 194677.73 389834.39 |

*Source: RStudio, Author’s work*

**Figure 18: Q-learning value functions by discount factors**

|  |
| --- |
| mdp\_Q\_learning (Pp, Rr, 0.95, N=10000)  $Q  [,1] [,2] [,3] [,4] [,5]  [1,] 103416.760 3747.782 797.2381 341.8786 970.3214  [2,] 194563.565 228118.981 396262.3713 197227.3056 199857.4366  [3,] 244235.641 204719.031 179800.9159 388439.6213 212070.0053  [4,] 5394.486 5840.316 6221.7503 147835.1031 8757.4350  [5,] 251981.994 267723.160 260166.1707 258515.1147 378751.7037  $V  [1] 103416.8 396262.4 388439.6 147835.1 378751.7  $policy  [1] 1 3 4 4 5 |

*Source: RStudio, Author’s work*

**Figure 19: Q-learning optimal policy for patient agents**

**C. Regression models**

|  |
| --- |
| Call:  lm(formula = log(Income90) ~ Collyr + TotExp + I(TotExp^2) +  RosenbSelfEst80 + OtisLenMat79, data = Occtest)  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 8.002315 0.478170 16.735 < 2e-16 \*\*\*  Collyr -0.069452 0.035379 -1.963 0.05662 .  TotExp 0.398349 0.147923 2.693 0.01030 \*  I(TotExp^2) -0.050590 0.019815 -2.553 0.01460 \*  RosenbSelfEst80 0.055805 0.019195 2.907 0.00592 \*\*  OtisLenMat79 0.003512 0.002036 1.725 0.09220 .  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.4148 on 40 degrees of freedom  Multiple R-squared: 0.3647, Adjusted R-squared: 0.2853  F-statistic: 4.593 on 5 and 40 DF, p-value: 0.0021 |

|  |
| --- |
| Call:  lm(formula = Collyr ~ RosenbSelfEst80 + OtisLenMat79, data = Occtest)  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -0.271065 2.035893 -0.133 0.895  RosenbSelfEst80 0.045176 0.085693 0.527 0.601  OtisLenMat79 0.006910 0.009031 0.765 0.448  Residual standard error: 1.865 on 43 degrees of freedom  Multiple R-squared: 0.0231, Adjusted R-squared: -0.02234  F-statistic: 0.5083 on 2 and 43 DF, p-value: 0.6051 |

|  |
| --- |
| Call:  lm(formula = log(Income90) ~ Collyr + TotExp + I(TotExp^2), data = Occtest)  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 9.54670 0.20997 45.466 <2e-16 \*\*\*  Collyr -0.04819 0.03912 -1.232 0.2248  TotExp 0.41220 0.16612 2.481 0.0172 \*  I(TotExp^2) -0.04997 0.02226 -2.245 0.0301 \*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 0.466 on 42 degrees of freedom  Multiple R-squared: 0.1579, Adjusted R-squared: 0.09774  F-statistic: 2.625 on 3 and 42 DF, p-value: 0.06288 |
|  |

*Source: RStudio, Author’s work*

**Figure 20: Various regression outputs for the income stochastic model**

**D. Cost function**

|  |
| --- |
| #Cost function  cf  [,1] [,2] [,3] [,4] [,5]  0 11790.44 23580.88 29476.1 47161.76  #Cost function with fully-funded PhD  cff  [,1] [,2] [,3] [,4] [,5]  0 11790.44 23580.88 29476.1 29476.1 |

|  |
| --- |
| #Initial ‘weighted-average’ evaluation  Bellman - cf[1,]  [1] 22053.192 10262.752 -1527.688 -7422.908 -25108.568  Bellman - cff[1,]  [1] 22053.192 10262.752 -1527.688 -7422.908 -7422.908 |

*Source: RStudio, Author’s work*

**Figure 21: Cost function by level of education with weighted-average education profile.**

**E. Synthetic model with private costs**

|  |
| --- |
| mdp\_Q\_learning(PIdt, exp(SLRr),0.95 )$V - cf[1,]  [1] 343158.61 308440.66 335472.61 105764.48 95541.07  disQler<-function(a){mdp\_Q\_learning(PIdt, exp(SLRr),a )$V - cf[1,]}  sapply(odisc, disQler)  #Value function net of costs  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 23620.144 24932.374 28000.879 32055.908 37398.561 44878.25 50856.389 74704.59 112158.826 220322.587 352289.0  [2,] -8906.783 -4672.177 -1466.199 2655.773 8151.735 10365.11 27387.284 46623.35 70924.785 192979.372 385751.0  [3,] -36342.062 -34992.177 -32815.870 -27664.227 -35633.260 -19091.28 -2935.901 16262.23 14228.998 165576.192 169620.2  [4,] -61567.858 -59128.817 -58899.331 -54365.621 -50793.225 -48758.93 -38290.042 -30731.59 -8480.602 68482.458 167561.9  [5,] -107047.858 -106257.191 -104379.331 -101964.950 -98745.781 -94238.93 -87478.668 -76219.93 -53681.687 -1813.124 48886.1 |

|  |
| --- |
| # Q-learning Modified optimal policy with reversion  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 2 2 1 1 2 1 1 2  [2,] 1 1 1 2 2 2 3 3 3 3 3  [3,] 1 1 1 2 2 2 2 3 3 3 4  [4,] 1 1 1 2 2 2 2 3 3 4 5  [5,] 1 1 1 2 2 2 2 3 3 4 5 |

*Source: RStudio, Author’s work*

Figure 22: Synthetic model of omitted transition and reward matrices with private costs

**F. Synthetic model with public costs**

|  |
| --- |
| disQler<-function(a){mdp\_Q\_learning(PIdt, exp(SLRr),a )$V - cf[1,]}  sapply(odisc, disQler)  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 23620.144 24932.3738 28048.920 32055.91 37398.559 44878.176 56097.439 74797.10 112118.02 223526.6 396696.1  [2,] 15033.217 16222.8405 19048.195 26595.70 32091.735 34305.112 44476.386 70544.27 93092.08 213859.3 339316.9  [3,] 11537.938 12887.8235 16093.801 20215.77 12246.756 33406.082 44947.555 64171.72 62226.68 126520.3 252684.1  [4,] -1717.858 -927.1833 2805.081 3365.05 9056.773 11091.070 17851.333 33568.27 51597.88 128130.0 236579.0  [5,] -11287.858 -10497.1833 -8619.331 -6204.95 -2985.775 1520.953 8280.126 19548.43 42062.62 101219.5 143809.4 |

|  |
| --- |
| # Q-learning modified optimal policy with reversion (self-funded)  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 2 2 1 1 2 1 1 2  [2,] 2 2 2 2 2 2 3 3 3 3 3  [3,] 3 3 3 3 3 4 3 4 3 3 4  [4,] 3 3 4 4 4 4 4 5 5 4 5  [5,] 3 3 4 4 4 5 5 5 5 5 5 |

|  |
| --- |
| disQler<-function(a){mdp\_Q\_learning(PIdt, exp(SLRr),a )$V - cff[1,]}    sapply(odisc, disQler)  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 23620.1433 24932.3738 28048.9206 32055.91 37398.558 44878.27 56097.83 74797.12 101469.64 194089.4 349516.7  [2,] 17917.9380 19267.8235 22473.8014 26595.77 32091.735 39786.07 51327.31 70553.41 109022.39 216458.0 307246.3  [3,] 11537.9380 12887.8235 16093.8014 20215.77 12246.775 33406.08 44947.60 37179.69 61105.96 207093.9 297560.8  [4,] -156.2474 -927.1833 2805.0812 3365.05 9056.775 11091.07 21559.72 34062.87 59018.14 131228.5 215426.8  [5,] -1717.8579 -927.1833 950.6687 3365.05 6584.219 11091.07 17845.75 29094.94 51589.88 116180.8 231286.6 |

|  |
| --- |
| # Q-learning modified optimal policy with reversion (fully-funded PhD)  # 0.05 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 0.95  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11]  [1,] 1 1 1 2 2 1 1 2 1 1 2  [2,] 2 2 2 2 2 2 3 3 3 3 3  [3,] 3 3 3 3 3 4 3 4 3 3 4  [4,] 3 3 5 5 5 4 4 5 5 4 5  [5,] 3 3 5 5 5 5 5 5 5 5 5 |

*Source: RStudio, Author’s work*

Figure 23: Value function and optimal policy by type with public education costs