A TEST FOR OVERCONFIDENCE WITH MARKET DATA

Daniel Ferrés*

Universidad de Montevideo
Department of Economics

RESUMEN

El fenómeno de excesivo optimismo (overconfidence) se refiere a creencias personales exageradamente optimistas sobre las propias habilidades individuales. El comportamiento excesivamente optimista de los agentes parece ser un patrón generalizado de las sociedades. Este elemento podría ayudar a explicar distintas decisiones irracionales desde el punto de vista económico. Recientemente, el optimismo excesivo ha sido sugerido en la literatura económica como un elemento que ayuda a entender la entrada exagerada de empresas a las industrias (y el alto porcentaje de quiebras en los primeros años de vida de las firmas que nacen) o el número ineficientemente alto de trades realizados en los mercados de capitales.

Por el momento, los economistas no han dedicado suficiente esfuerzo para documentar evidencia sobre el comportamiento excesivamente optimista de los individuos. La mayoría de la información sobre el tema surge de entrevistas, encuestas o experimentos. Si bien estos métodos tienen cierta validez, también es verdad que reciben fuertes críticas. En particular, no hay consenso entre los economistas sobre la robustez de la información generada a partir de experimentos realizados en laboratorios. Quiere decir que los economistas no se han tomado el fenómeno del optimismo excesivo tan en serio como los psicólogos y economistas del comportamiento sugieren que deberían hacerlo.

En esta primera versión de mi trabajo, presento una prueba de optimismo excesivo de los agentes basada en observaciones de la realidad. En particular realicé un test en base al comportamiento individual a la hora de buscar trabajo basado en datos de NLSY (del Departamento de Trabajo de Estados Unidos). Esta versión se trata de una estimación preliminar del test. Los resultados obtenidos son consistentes con la presencia de excesivo optimismo en las personas. En una etapa posterior del trabajo, utilizaré una metodología que me permitirá obtener resultados más robustos.

*I thank Juan Dubra for his helpful comments. All errors are of my own responsibility.

ABSTRACT

Overconfidence refers to the exaggerated beliefs that one might have about his own personal abilities. This phenomenon seems to be pervasive and it could help explaining several economic paradoxes. Still the profession has failed to produce conclusive evidence about overconfident behavior using market data. Most of the evidence that shows that people are overconfident comes from either psychological interviews, surveys or from experiments in economic laboratories. Although these method doubtless have their validity they are not free of criticism, and therefore economists probably do not take overconfidence as seriously as psychologists and behavioral economists suggest they should.

I present a working paper version of an empirical test for overconfident behavior using data about individual conduct when searching for a job. Results are consistent with the presence of overconfident behavior. Note that this work constitutes only a first approach to the test. In particular, I will improve methodological aspects of my paper in a future version of my work so that I can get more robust results.

A Test for Overconfidence with Market Data

1. Introduction

Psychological studies show that most people are overconfident about their own relative abilities, and unreasonably optimistic about their futures. For example, 90 per cent of American drivers in one study thought they ranked in the top half of their demographic group in driving skill (Ola Svenson, 1981). Other traits on which people have been shown to overestimate the chances of good events are income prospects and longevity.

The belief among behavioral economists that overconfidence is pervasive has produced two growing strands of literature. The first is aimed at studying its economic consequences. The pioneering works of Manove (1995) and Manove and Padilla (1998) have shown that overconfidence on the part of some entrepreneurs leads to a screening problem by banks and to inefficiencies in the allocation of credit. Dubra (2004) has shown that overconfidence on the part of the unemployed leads to longer search spells. The second branch of the behavioral literature on overconfidence is aimed at showing that this bias can explain several paradoxes. For example, it has been observed that overconfidence can explain why only a minority of new firms are alive five years after opening and Odean (1999) argues that overconfidence is probably the reason why trading volume is too high.

Although overconfidence seems to be pervasive, and it could help explaining several economic paradoxes, the profession has failed to produce a conclusive proof that overconfidence is pervasive using market data. Most of the evidence that shows that people are overconfident comes from

---

1 Timothy Dunne et al (1988) using plant-level data from the US Census of Manufacturers spanning 1963-1982 estimated that 61.5% of all entrants exited within 5 years and 79.6% within ten years. The idea that overconfidence causes business entry mistakes has been explored by Richard Roll (1986) and Camerer and Lovallo (1999).
either psychological interviews or surveys, or from experiments in economic laboratories. Although these methods doubtless have their validity they are not free of criticism, and therefore economists probably do not take overconfidence as seriously as psychologists and behavioral economists suggest they should. One of the most usual criticisms of experimental data is that it is produced with people who are either not experts in the decisions they make (i.e. undergrads) or face insufficient incentives (usually low prizes), or that when experts are used in experiments involving the tasks they usually perform, with appropriate incentives, they may still take the experiment as just a game, not involving their reputation or job. This view that experiments are a weak tool for economists is still alive, and has had several advocates. The validity of some papers has been questioned on the grounds that their main, or only, hypothesis comes from experimental data (see Palacios Huerta, Serrano and Volij on Rabin).

In this paper, I will try to prove, using market data, that overconfidence is indeed pervasive. This is important, for three reasons. First, it will validate the findings of dozens of psychological and experimental studies that show the existence of overconfidence. Second, it will also validate the empirical relevance of the theoretical papers that use overconfidence as one of their hypothesis. Finally, it will help our understanding of economics, since it will validate the explanations of paradoxes that are based on overconfidence.

2. The model

In order to explain how my test works, I will first present a simple example of a search model in which the searcher is overconfident, and show how overconfidence affects his behavior. The example is based on the more general model by Dubra (2004).

Suppose that an unemployed worker will receive in each future date a wage offer that can be either $1 or $2. The true distribution, $f$, from which the offers are drawn is $f(1) = f(2) = \frac{1}{2}$. In standard models, the searcher knows this distribution. Let us assume, however, that the searcher does not know the distribution, but that he believes that the offers are drawn from one of two urns. The first urn is $g$, where $g(1) = \frac{3}{4} = 1 - g(2)$ and the second is $h$, with $h(1) = \frac{1}{4} = 1 - h(2)$.

We will call the searcher unbiased if he believes that the urn is $g$ or $h$ each with probability $\frac{1}{2}$ and overconfident if he believes that the urn is $h$ with probability larger than $\frac{1}{2}$. An unbiased searcher is correct, on average, about his chances of getting an offer of $1$ or $2$, since

$$P(1 \mid \text{unbiased}) = P(g) \cdot g(1) + P(h) \cdot h(1)$$
$$= \frac{1}{2} g(1) + \frac{1}{2} h(1) = \frac{1}{2} = f(1).$$

The overconfident assigns a higher probability than the truth to an offer of $2$, however, since for a belief $p > \frac{1}{2}$ that the urn is $h$ we obtain

$$P(1 \mid \text{overconfident}) = P(g) \cdot g(1) + P(h) \cdot h(1)$$
$$= (1-p) g(1) + p h(1) < \frac{1}{2} = f(1).$$

2 The fact that the searcher does not consider the true urn $f$ to be a possibility is not important in this example.
The key feature of this simple example is that while after receiving one wage offer the unbiased remains unbiased, after receiving one offer, the overconfident becomes less overconfident: his beliefs decrease. I will now show this feature with a calculation.

The posterior belief of an urn of $h$ for the unbiased after an offer of 1 is

$$P(h \mid \text{past offer of } 1) = P(1 \mid h) P(h) / P(1) = \frac{1}{2} \cdot \frac{1}{4} / \left[ \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{3}{4} \right] = \frac{1}{4}$$

And similarly, $P(g \mid po = 1) = \frac{3}{4}$. Therefore, after an offer of 1, the unbiased searcher believes that the chance of an offer of $1$ is:

$$P(1 \mid po = 1) = P(g \mid po = 1) P(1 \mid g) + P(h \mid po = 1) P(1 \mid h)$$

$$= \frac{3}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} = 5/8.$$ 

Similarly, his posterior after an offer of 2 is:

$$P(h \mid po = 2) = \frac{1}{4}$$

and $P(g \mid po = 2) = \frac{1}{4}$, so that after an offer of 2, the unbiased searcher believes that the chance of an offer of $1$ is:

$$P(1 \mid po = 2) = P(g \mid po = 2) P(1 \mid g) + P(h \mid po = 2) P(1 \mid h)$$

$$= \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} = 3/8.$$ 

Given these posteriors, the modeler knows that after one offer, the posterior beliefs of an unbiased searcher that he will receive an offer of $1$ is

$$P(1 \text{ after one offer}) = f(1) P(1 \mid po = 1) + f(2) P(1 \mid po = 2)$$

$$= f(1) \cdot \frac{5}{8} + f(2) \cdot \frac{3}{8} = \frac{1}{2}.$$ 

That is, the modeler knows that after one (or more offers) the unbiased searcher will remain unbiased on average. This sounds like the usual claim that “beliefs are a martingale” but it is different, however. For the calculation of “beliefs are a martingale” we should have used in the last equation not $f$, but the subjective beliefs. In this case they coincide only because the searcher is unbiased, and this is precisely why the unbiased searcher’s beliefs are a martingale. For the overconfident, when we average his posteriors tomorrow, using the true distribution, we find that he tends to become less overconfident, as I now show.

The posterior belief of an urn of $h$ for the overconfident after an offer of 1 is

$$P(h \mid po = 1) = P(1 \mid h) P(h) / P(1) = p \cdot \frac{1}{4} / \left[ p \cdot \frac{1}{4} + (1-p) \cdot \frac{3}{4} \right] = p / (3 - 2p)$$

And similarly, $P(g \mid po = 1) = 3(1-p) / (3-2p)$. Therefore, after an offer of 1, the overconfident
searcher believes that the chance of an offer of $1 is:

\[ P(1 \mid \text{po} = 1) = P(g \mid \text{po} = 1) \, P(1 \mid g) + P(h \mid \text{po} = 1) \, P(1 \mid h) \]
\[ = (3(1-p) \, 3 + p) / 4(3 - 2p) = (9 - 8p) / 4(3 - 2p) \]

Posteriors after an offer of 2 are:

\[ P(h \mid \text{po} = 2) = P(2 \mid h) \, P(h) / P(2) = p / [ p / 4 + (1-p) / 4 ] = 3p / (1 + 2p) \]

and \( P(g / \text{po} = 2) = (1-p) / (3 + 2p) \). We obtain that the subjective probability that the overconfident assigns to receiving an offer of 1, after he has received an offer of 2 is:

\[ P(1 \mid \text{po} = 2) = P(g \mid \text{po} = 2) \, P(1 \mid g) + P(h \mid \text{po} = 2) \, P(1 \mid h) \]
\[ = ((1-p) \, 3 + 3p) / 4(1 + 2p) = 3 / 4(1 + 2p) \]

I am now able to show what I wanted: an overconfident searcher \((p > ½)\) will assign on average a higher probability to an offer of 1 after he has received one draw. That is,

\[ P(1 \text{ after one offer}) = f(1) \, P(1 \mid \text{po} = 1) + f(2) \, P(1 \mid \text{po} = 2) \]
\[ = (9 - 8p) / 8(3 - 2p) + 3 / 8(1 + 2p) \]

and

\[ P(1 \text{ after no offers}) = 3(1-p) / 4 + p / 4 \]

One can then show that \( P(1 \text{ after one offer}) > P(1 \text{ after no offers}) \)

Therefore, the calculation shows that on average, according to the true distribution, the overconfident will become more and more pessimistic.

I have shown in this simple example that unbiased individuals tend to remain unbiased over time, and that overconfident individuals tend to become more and more pessimistic. This observation joined with the rather obvious feature of search models that as individuals become more pessimistic, they will search less and less (if the true distribution of offers remains constant) yields the prediction of behavior that I will test:

“If people are unbiased, as more and more information arrives, their search spells will remain constant. On the contrary, if individuals are overconfident, in the first spell they will receive relatively bad information, and will thus become more pessimistic about their abilities, and that will make them search less in their future spells. Similarly, the information gathered in their second spell, will make overconfident individuals sample less in their third spell, and so on and so forth”.

\(^{3}\) I used information about the duration of unemployment spell. But I think that it would be more appropriate to measure the duration of the job search. I will further discuss this issue in section 3.1.5.
That is: if I show that search spells tend to get shorter, I will have found evidence that is consistent with individuals being overconfident.

3. Data Set

I use information from The National Longitudinal Survey of Youth 1979 (NLSY79). The National Longitudinal Surveys (NLS) are a set of surveys sponsored by the Bureau of Labor Statistics, U.S. Department of Labor. These surveys have gathered information at multiple points in time on the labor market experiences of diverse groups of men and women. NLSY79 is a nationally representative sample of 12,686 young men and women who were 14-22 years old when they were first surveyed in 1979. As said, NLSY79 gathers information from sample members about their work, family, investments on education and other life course experiences. Data has been collected annually from 1979 to 1994 and biennially from 1996 to 2000.

NLSY79 sample contains three sub samples:
- a sample of 6,111 respondents that represent the non-institutionalized civilian segment of people living in the U.S. in 1979 that were 14-21 years old as of December 31, 1978.
- a supplemental sample of 5,295 respondents designed to over sample civilian Hispanic, black, and economically disadvantaged non-black/non-Hispanic people living in the U.S. in 1979 that were 14-21 years old as of December 31, 1978.
- a sample of 1,280 respondents designed to represent the population that was 17-21 years old as of December 31, 1978 and who were enlisted in one branch of the military as of September 30.

3.1. Variables of interest

I am interested in the relationship between a variable called ‘Duration of Unemployment Spell’ and the ‘Number of Unemployment Spell’. I think that an important part of my paper is controlling for the many factors that are likely to affect the duration of an unemployment spell. By doing this I am trying to compute the “pure” belief effect. It is easy to list variables that could be included in my regressor vector. I started my paper controlling for ‘Age’, ‘Economic Resources Available’ and ‘Number of Dependants’. This means that I recognize the possible effect of these three independent terms on the y variable. Still I know that this specification may be improved in order to lessen the omitted variables bias (I will further comment on this issue later).

In this section, I will mention some facts about the data available on the NLSY79 on the variables that I am considering. Also I present the reasons why I will restrict the selection of my observations to the information collected during the rounds conducted between 1994 and 2000.

3.1.1. Some comments about data regarding the ‘Age’ variable

I am interested in Age as a factor affecting behavior when searching for a job. Because NLSY79’s respondents are 14-22 years old when the survey started, I did not take into account the first rounds of the survey (many questions are restricted to respondents being over, 16 or over 18).

3.1.2. Some comments about data regarding ‘Economic Resources Available’
Income and wealth are both measures of Economic Resources Available to a respondent. I think that it is important to control for this factor, since it indicates the capacity to wait for an “appropriate offer” while he is searching for a job. Wealth is equal to an individual’s total assets minus his/her total debts. I consider that wealth (or net wealth) is a better indicator of an individual’s financial position than income.

To protect confidentiality the survey used top coding for unusually high values, both for questions about wealth and income. The NLSY79 has used different top coding algorithms for assets and income questions in different periods of time. Beginning in 1996, the top two percent of respondents with valid values were identified. Values within that top range were averaged and that averaged value replaced all values in the top range. The extent of top coding for NLSY79 asset and income questions varies greatly. For example, in 1993 there were only two individuals whose money assets exceeded the cut-off value of $500,000, while 581 individuals gave a market value for their residence above the cut-off value of $150,001.

One major concern when asking individuals about their asset holdings is nonresponse bias. In my case I am interested in individual’s specification about the value of their holdings. Factors that are likely to contribute to nonresponse are suspicion, uncertainty about an asset’s current value, shared responsibility for family finances, and complex financial arrangements. Problems arise also as questions refer to the household’s asset holding or income. Respondents are often certain about the value of their asset holdings but know their spouse’s wealth (or income) with less certainty and partner’s wealth (or income) with much less certainty.

Between 1979 and 1982, NLSY79 respondents were only asked questions on assets and income if they met one of the following five criteria: (1) 18 years old or greater; (2) had a child; (3) enrolled in college; (4) married; (5) living outside their parent’s home. This selection process eliminated many respondents from these questions. For example, in 1979 only five percent of those interviewed under age 18 answered the assets questions. Also consider that ‘Asset’ questions were dropped during 1983 and 1984. Beginning in 1985, when all respondents had turned 18, NLSY79 respondents were administered a large range of questions about wealth (about 20 questions about asset and debt holdings). As the cohort has aged, the wealth section has grown in length and detail, so that the survey provides a rough overview of the net worth of each respondent. Budgetary restrictions resulted in the elimination of wealth questions for the surveys conducted in 1991 and 2002.

3.1.3. Some comments about data regarding “Number of Dependents”

Many variables may be considered as a proxy to ‘Number of Dependents’. I decided to use the variable ‘Number of Own Children in Household’ (it is a created variable reporting number of biological, step and adopted children in household). ‘Household’ is defined as “all individuals sharing the respondent’s

---

4 Total Assets is the addition of the value of an individual’s residential properties, vehicles, business, financial assets (CDs, IRAs, Money Market accounts, etc) and a category called “other assets”. Total Debts is the sum of an individual’s debt by concept of mortgage, other debt on property, debt on business, debt on vehicles and a category called “other debts”.

5 Income questions report earnings from every source in the previous calendar year.
primary residence at the time of interview”. Also it is a good idea to include a dummy variable that takes a value of 1 if the spouse or partner is unemployed at the moment of the survey.

3.1.4. Some comments about data regarding ‘Number of Unemployment Spell’

In each round of the survey, respondents where asked to report the length of each unemployment spell since the last interview. I constructed the variable “Number of Unemployment Spell” by counting the number of times when a respondent reported the length of the period of unemployment. I start counting from the 1994 survey. Every round asks for the duration of at most four periods of unemployment. Suppose that the respondent was unemployed one time between the 1993 survey and the 1994 survey; two times between the 1994 survey and the 1996 survey; he was never unemployed between the 1996 survey and the 1998 survey; and he had one unemployment spell between 1998 and 2000. Then the last unemployment spell will be unemployment spell number four. So counting the number of unemployment spell is possible.

3.1.5. Some comments about data regarding ‘Duration of Unemployment Spell’

As said, individuals self-report the duration of each unemployment spell. Duration is measured in weeks. Note that I am interested in the duration of the unemployment spell. It could be argued that I should consider the duration of the job search instead of the period of unemployment. An individual who is unemployed, might be out of the labor force too. This implies that he is not looking for a job and therefore he does not receive job offers (that inform him about what a potential employer thinks about his talent). So he is not actually “updating his beliefs”. This different approach will allow me to make this interesting distinction.

Also, I think that information about methods of searching might be relevant. I did not account for this issue in my study but it might be worthwhile to control for search intensity in a later version. My intuition is that greater searching intensity lowers the expected duration of the unemployment spell.

3.1.6. Some final comments about the Data Set

The 1987 round was mostly done by telephone interviews, so I decided not to include this particular year in my study. By saying this, I am stating that the different method used when conducting the survey, may affect the answers.

I conclude this section by saying that after considering all the specific points commented above I have decided to restrict my attention to observations collected in the rounds conducted in 1994, 1996, 1998, 2000. Naturally, this decision also has the advantage that I will be working with more updated data.

---

6 I explain later why I could not include this dummy variable.
7 Note that in 1994 the NLSY started to be conducted every two years, instead of every year. This means that in the 1994 round I have information about unemployment spells corresponding to a one year period (from the 1993 survey to the 1994 survey) while in the other rounds I have information about all the unemployment spells that occurred in a two-year interval. I think that the only consequence of this is that I have a slight downward bias in the number of unemployment periods. Still, I believe that this issue is not so much important.
3.2. Summary statistics

I present in this section some summary statistics about the sample:

Table 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td># of unemployment periods (per person)</td>
<td>0.021</td>
<td>0.204</td>
<td>0.00</td>
<td>8.00</td>
</tr>
<tr>
<td>duration of unemployment period</td>
<td>39.61</td>
<td>40.98</td>
<td>0.00</td>
<td>744.00</td>
</tr>
<tr>
<td>children in HH</td>
<td>1.40</td>
<td>1.31</td>
<td>0.00</td>
<td>9.00</td>
</tr>
<tr>
<td>age</td>
<td>35.82</td>
<td>3.17</td>
<td>29.00</td>
<td>44.00</td>
</tr>
<tr>
<td>total assets</td>
<td>74,176.74</td>
<td>194,727.80</td>
<td>0.00</td>
<td>6,715,768.00</td>
</tr>
<tr>
<td>net assets</td>
<td>18,442.99</td>
<td>237,726.10</td>
<td>-3,990,233.00</td>
<td>5,812,767.00</td>
</tr>
</tbody>
</table>

The table presented below refers to the maximum, minimum and average duration of each number of periods of unemployment. “Observations” should be thought of as the number of respondents that had a particular number of periods of unemployment. For example, for 49 respondents, 4 was the total number of periods of unemployment between the 1993 interview and the 2000 interview.

Table 2

<table>
<thead>
<tr>
<th>wu1</th>
<th>wu2</th>
<th>wu3</th>
<th>wu4</th>
<th>wu5</th>
<th>wu6</th>
<th>wu7</th>
<th>wu8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max</td>
<td>744.0</td>
<td>135.0</td>
<td>131.0</td>
<td>118.0</td>
<td>108.0</td>
<td>52.0</td>
<td>52.0</td>
</tr>
<tr>
<td>Min</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>5.0</td>
</tr>
<tr>
<td>Average</td>
<td>41.7</td>
<td>39.0</td>
<td>33.6</td>
<td>37.1</td>
<td>23.0</td>
<td>17.7</td>
<td>29.3</td>
</tr>
</tbody>
</table>

I am aware that 744 sounds like too big a number for the duration of an unemployment spell. There are 12 observations for which the duration of the first period of unemployment is higher than 200 weeks. I think that they are strange. I initially run two regressions: one where I considered the series including the observations above 200 weeks and I eliminated later the 12 observations above 200. Coefficients and significance levels where very similar. So I decided not to censor the series and keep working with all the observations.

A quick note on notation: wu1 designates period of unemployment number 1, wu2 indicates period of unemployment number 2, and so on...
4. Estimation and Results

I estimate the parameters of interest using Ordinary Least Squares (OLS). Before presenting the results, I want to comment a methodological aspect of my present work. As measurement of “duration of unemployment spell” (the dependent variable) is made while unemployment is ongoing, data may present a censoring problem. For individuals that were unemployed at the time they were surveyed, duration is at least the reported time, but not necessarily equal to it. OLS estimation is not an entirely sound methodology to use in these cases, because it may lead to biased and inconsistent estimates. In the case of this particular data set, censoring does not look like a significant issue. At this stage of my work I present preliminary results because I find that OLS estimation leads to an intuitive and handy test. In any case, in a future version of my work, I will use Hazard Rates Models, where assumptions about the distribution of unemployment duration are required. Hazard Rates Models eliminate a potential censoring problem.

I now emphasize the relationship of interest. If I verify a negative (and statistically significant) relation between duration and number of unemployment spell, then there is room to think that we found overconfident behavior. I do not mean that if this relationship was confirmed, then it is sufficient evidence to say that overconfidence is present in this sample. Only after controlling for other factors affecting the dependent variable I would be in a position to say that the “pure belief effect” causes the negative duration-number of unemployment correlation.

To test for this negative relationship I will use different specifications. In all cases I regress the logarithm of the variable indicating the duration of the unemployment spell (logwu) on the logarithm of the unemployed’s age (logage), the logarithm of his total assets (logtotalassets) and a proxy indicating the number of dependents (logchi)\(^{10}\). Also I will include in my regressor vector the number of period of unemployment (nrwu). This variable will enter my regression in different forms:

- I allow for the number of the unemployment spell (nrwu) to enter the regression (as in model M1 in table 3);
- I include the log of nrwu (in model M2);
- I introduce dummy variables indicating the number of the period of unemployment (in model M3);
- I present a quadratic specification (model M4).

These different approaches will permit me to make a somewhat significant remark. In the M1 and M2, the marginal effect on the duration of the unemployment period of having one more period of unemployment is constant. In the cases where I include dummy variables and in the quadratic

---

\(^{10}\) Employment status of respondent’s spouse is relevant when assessing the number of dependents. I found in the NLSY some general questions about whether respondent’s wife was employed or not at the moment of the interview. Still answers to these questions did not help specifying whether respondent’s spouse was employed at the moment when the interview took place. For this reason, I judged that I did not have good information about whether the spouse should be considered a dependent or not. I decided not to include spouse as dependents.
specification, I am allowing for different marginal effects as the individual moves from their first to their second period of unemployment, from the second the third one, etc.\textsuperscript{11}

In models M5, M6 and M7 I add two dummy variables indicating “Region” and “Race”. “Region” takes a value of 1 if the respondent is in “urban” residence and 0 if he is in “rural” residence. “Race” takes the value of 1 if the unemployed is “white” and 0 if he is “black or other”\textsuperscript{12}. Dummy variables specify that we are dealing with different types of individuals. They help recognizing that unemployment spell duration may differ across individuals just because one is white and the other one is non-white. The reason why I include these binary variables is that I am trying to improve the specification of the model. As I commented earlier, I have to control for every variable affecting the dependent variable in order to get the ‘pure belief effect’. Controlling for every factor is impossible. I just included this two as an initial step.

\textsuperscript{11} Actually I only created three dummies that indicate whether it is the first, second and third or above period of unemployment. The reason for this is that there are not so many observations for the fourth, fifth, sixth, seventh or eighth period of unemployment. Still, the idea of the different marginal effects as unemployment periods arrive is present.

\textsuperscript{12} For some individuals I found valid answers in some years, no interview in some other years or a valid skip. In this cases I used the following criteria: if a respondent states that he is white at least one time and there is no answer for the rest of the rounds I count him as white. In those cases where an individual actually responds a “Race” giving different answers in different years, I count him as whatever he responded more times.
In the following part of this section I will present the empirical results and interpret the estimates. Estimation of alternative regressions are presented in Table 3.

<table>
<thead>
<tr>
<th>Table 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>logwu</td>
</tr>
<tr>
<td>constant</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>lognrwu</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>lognrwu2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>nrwu</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>logtotalassets</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>logchi</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>logage</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>region</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>race</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Du2</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Duover3</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Obs</td>
</tr>
<tr>
<td>R2</td>
</tr>
<tr>
<td>BIC</td>
</tr>
</tbody>
</table>

* The coefficient is significant at 1% level.
** The coefficient is significant at 5% level.
*** The coefficient is significant at 10% level.
In model M1, I regress the logarithm of the variable indicating the duration of the unemployment spell (logwu) on the number of period of unemployment (nrwu), the logarithm of the unemployed’s age (logage), the logarithm of total assets (logtotalassets) and a proxy indicating the number of dependents (logchi). I find that this specification is very handy because of its straightforward interpretation: for the independent variables that enter the regression in the log form, coefficients are elasticities. As may be observed in table 3, a 1% increase in age causes a 1.55% increase in average duration of the unemployment spell. The interpretation for nrwu goes as follows: as one more period of unemployment arrives, its expected duration decreases by 17.04% (holding everything else constant). In model M2, lognrwu enters the regression instead of the actual number of the period of unemployment. The interpretation is not as neat as in M1, but it is important to emphasize that the estimated coefficient for both variables (in both models) are negative and statistically significant at 1% level.

I introduce dummy variables to indicate the number of the unemployment period in model M3. Holding everything else constant, the difference in logwu as an individual arrives to her second period of unemployment is -.1547. This means that the expected duration of this unemployment spell is 15.47% lower than if it was her first period of unemployment. In model M4 I suggest a quadratic specification. Under this quadratic form, the marginal effect of going from one unemployment spell is not constant. Note that if:

$$\log wu = \alpha + \beta_1 \log nrwu + \beta_2 \log nrwu^2 + \beta_3 \log chi + \beta_4 \log total assets + \beta_5 \log age + \mu$$

then:

$$\frac{\partial \log wu}{\partial \log nrwu} = \beta_1 + 2 \cdot \beta_2 \log nrwu \quad (1)$$

In particular, the derivative presented in (1) is decreasing in nrwu (it is less negative as nrwu increases).

Models M5, M6 and M7 are alternative regressions to the ones presented earlier. As said, I introduce dummy variables for “Race” and “Region”. In these regressions I have combined different use of the variables nrwu log nrwu and the Dummy variables Du2 and Duover3. The negative relationship of interest is confirmed in these three models. In fact, the estimates of the coefficients for nrwu and log nrwu in models M5 and M6, respectively, are almost unchanged when compared with their estimates in models M1 and M2. Note that I could not reject the null hypothesis that the coefficient for the variable “Race” (when it takes a value of one) equals zero in models M5 and M6 at a 10% significance level.

I will focus now on the discussion of the coefficients of logage, logchi and logtotalassets. I will comment their coefficients and the rejection of the null hypothesis of the t-test. Coefficients for

---

13 I also run these regressions putting lognetassets in the regressor vector instead of logtotalassets. AIC and BIC values for these alternative regressions were generally higher. Also estimates were very similar but some of these coefficients were not statistically significant.
each variable are of the same sign (and similar magnitude) in the seven models presented. Age and number of dependents appear to be positively correlated to the duration of the unemployment spell\textsuperscript{14}. For unemployed’s total assets the inverse relation holds\textsuperscript{15}. I already commented about how this coefficients should be interpreted (they are elasticities). For each of these three variables, the null hypothesis that individual coefficients are zero is rejected generally at the 1% significance level (indicated with asterisks in the table). This allows me to state that age, number of dependents and total assets are relevant to explain variations in the dependent variable.

I also report in Table 3 the value for Akaike Information Criterion (AIC) and Schwartz Information Criterion (BIC). Both are popular methods of model selection. The rule is to select a model among a group of models if it has the lowest AIC or BIC. When comparing this seven alternative models, M5 and M6 are the two models that have lower AIC and BIC. It could be argued that both specifications include at least one variable that is statistically insignificant. In general, I would feel comfortable with a model that on the one hand introduces a variable that causes some other to be statistically insignificant but on the other hand leads to a significantly lower value of both AIC and BIC.

Finally, I provide information about the $R^2$ of each regression. Values are very low for the seven models presented. Even though I would like to see a higher value of $R^2$, I am not particularly concerned about it\textsuperscript{16}. I have not used a structural model to estimate duration of the unemployment spell. I believe that if I would have used a structural model, $R^2$ would have been higher. Also, I know that there are many omitted variables that could be included in order to try to increase the explanatory power of this model (as I included the binary variables for “Race” and “Region”). But that is not the main purpose of my paper. I am interested in finding possible evidence for the negative relationship between duration and number of unemployment spell. If I look at Table 3, it is clear that the intuitive relationship that I am testing is robust to the many specifications of regressions presented in my study. Anyway, in later versions of my work I will control for aspects like Education, Gender and a proxy for the Economic Situation at the time of search (not personal, but at the aggregate level).

5. Conclusion

I suggested a test for overconfidence based on market behavior data. Most of the evidence that shows that people are overconfident comes from either psychological interviews, surveys, or from experiments in economic laboratories. My test is based on a model that suggests that a negative relationship between duration of unemployment spells and the number of the period of unemployment is evidence for overconfident behavior.

\textsuperscript{14} I admit that initially I anticipated a negative sign for the coefficient of logchi.

\textsuperscript{15} One could think that if the individual is richer, he would probably got a higher level of education (and his skills are better). This could explain why his expected unemployment spells are shorter than those of an individual who is not as rich.

\textsuperscript{16} In this type of micro econometrics models, it is not unusual to get these values of $R^2$. 
I have examined data available on NLSY about unemployed’s behavior as they search for a job. After controlling for some of the variables affecting the duration of an unemployment spell, I found evidence supporting the hypothesis that duration and the number of unemployment spell are negatively related. This relationship has proved to be robust to the seven different specifications that I present in Table 3. This finding could be interpreted as overconfident behavior of individuals.

However, as stated by Dubra (2004), there are various alternative theories that could explain why search spells get shorter. First, it could be the case that there is “learning by doing” on the part of searchers: they learn to search, and thus find an appropriate job in less time. A second theory that could explain shortening spells is Berkovitz’s stigma theory of unemployment. Berkovitz suggested that when a firm hires a worker, it learns its type, and that when a firm learns that the worker is of a low type, it is more likely to fire him. Therefore, a large number of spells indicates to the firms in the market that the worker is of a low type. This implies that the distribution of wage offers will decrease in first order stochastic sense, with each additional unemployment spell. This in turn makes searching not a worthwhile activity, and search spells get shorter over time. So, before concluding that spells get shorter because people are overconfident, I would need to control for these two aspects impacting on the evolution of unemployment spell duration.

Last, I will comment my short term strategy in order to improve my work. First, I will use Hazard Rate Models in order to prevent censoring problems. Second, I will incorporate a few new explanatory variables in order to increase the explanatory power of the model. Clearly, I need to control for “Education” and “Gender”. Third, I will follow Dubra’s (2004) suggestions in order to control for the two phenomena commented in the above paragraph (“learning by doing” and “stigma”).

6. References


