A Replication of "The importance of capital in closing the entrepreneurial gender gap: A longitudinal study of lottery wins"

ryooc01

July 2024

1 Introduction

In the realm of economic research and policy endeavors, understanding the factors that influence entrepreneurship remains an open question, particularly when considering gender disparities in the field. The study "The Importance of Capital in Closing the Entrepreneurial Gender Gap: A Longitudinal Study of Lottery Wins" by Sarah Flèche, Anthony Lepinteur, and Nattavudh Powdthavee, published in the Journal of Economic Behavior and Organization, aims to provide some insight into the gender entrepreneurial gap by leveraging longitudinal data on lottery winners. The authors demonstrate that large lottery wins (defined as lottery wins within the top 25th percentile of survey respondents, *i.e.*, lottery wins greater than or equal to £86.66) are correlated with a roughly 2 percentage point increase in the probability of being self-employed. The authors use their findings to conclude that capital accessibility is a pivotal factor in reducing gender disparities in entrepreneurship.

In the sections that follow, I will first outline the key specifications and underlying assumptions of the original study. Next, I will detail the methodology employed for replication, followed by a comparison of the summary statistics, figures, and tables between the original findings and my replicated results.

2 Paper Background and Assumptions

Fleche et al. use data from the British Household Panel Survey (BHPS), a nationally representative sample of over 10,000 households and nearly 30,000 individuals. Although the authors do not explicitly specify the years of data they use in their analysis, they state that the BHPS was conducted between September and Christmas of each year from 1991 to 2008, and data collection on lottery winners began in 1997. Based on the number of observations reported in Table 1, we conclude that Fleche et al. must have used all 18 waves of data in their analysis.

The authors limit all analysis to working-age adults, defined as individuals aged 16 to 65. They also adjust the prices of lottery winnings and income for inflation, using CPI data with 2000 as the base year.

The BHPS dataset is organized into a series of .dta files for each wave. The two key files relevant to this paper's analysis are *indresp* and *hhresp*. Each wave is indicated by a corresponding letter as a prefix in the file name, representing the wave number. For instance, the files for wave one are named aindresp and ahhresp, while those for wave two are bindresp and bhhresp, continuing in this pattern for subsequent waves.

In order to link variables within waves, there is a *person number* (denoted by the root 'pno') and a *household identification number* (denoted by the root 'hid'). Using these two IDs, one can obtain a unique ID for each individual within a wave. This combination of household ID and person number is used to link variables across files. It is important to note that these ID numbers are not consistent across waves. So, in order to link individual information across waves, there is another variable called the *cross-wave person identifier* (denoted by the variable name 'pidp'). Note that this variable does not take on the wave-letter prefix; it is standardized across waves.

3 Overview of Replication Methodology

I began by loading the necessary libraries into R (margins, tidyverse, haven, dplyr, lme4, and stargazer). I created a list of the relevant columns for this analysis and initialized another list to store the intermediate data sets of each wave:

```
columns <- c("hid", "pno", "sex", "jbstat", "doby", "windfg", "windfgy",
"mastat", "pid", "pidp", "fihhyr", "hlstat", "hlsf1", "hospd", "nkids",
"qfedhi", "qfvoc", "region.x")
mdfs_list <- list()
```

The list of column names only contains the "root" column names that are consistent across all waves, so in order to obtain the wave-specific files and variables, I used a for loop to cycle through the files in each wave, then upload and identify the necessary columns. After retrieving the files within the wave, I used the household identification number and the person number to merge the necessary files within a wave to get an aggregated wave data set that contains all the relevant variables, called merged_df. I also added a variable that indicates wave number. I then selected the columns indicated in my list of desired columns (shown above).

```
final_df <- NULL

# Select available columns
available_columns <- intersect(paste0(letter, columns), colnames(merged_df))
```

```
if (length(available_columns) > 0) {
    final_df <- select(merged_df, all_of(available_columns), pid, pidp, wave)
} else {
    message(paste("Columns not found in data frames for letter:", letter))
    next
  }
# Rename columns to remove the prefix letters and keep variable names consistent
colnames(final_df) <- gsub(paste0("^", letter), "", colnames(final_df))
```

I then merged all of the cleaned intermediate data sets to get one long master data frame containing data on all the individuals in the BHPS data set. The final cleaned data frame of working-age individuals contained 238,996 observations. After restricting the data set to working-age individuals (calculated by subtracting the date of birth from wave year), I was left with 192,543 observations.

I then adjusted prices of lottery winnings and earnings using CPI data from the Office of National Statistics, as indicated by Fleche et al. (link: https://www.ons.gov.uk/economy/inflationandpriceindices/timeseries/d7bt/mm23). Note that this CPI data has a base year of 2015 and Fleche et al. use a base year of 2000, so adjustment is needed. After these steps, the data set was sufficiently cleaned to begin replicating summary statistics and analysis.

4 Comparative Summary Statistics

For this section, I will focus on areas of inconsistency between Fleche et al. and my replication, and areas that I believe warrant any additional explanation. See Figure 1 for a list of summary statistics found in Fleche et al. and my own replication.

Fleche et al. report that 36% of working-age BHPS individuals report at least one lottery win. When replicating this finding, I found that only 25% of working-age BHPS individuals reported at least one lottery win. When restricting the data to the years when lottery data was available (1997 onwards), the percentage of individuals who reported at least one lottery win increased to only 26.23%.

```
#Calculating the percentage of individuals who report at least one lottery win:
winners <- df %>%
  filter(windfg == 1) %>%
  filter(windfgy > 0)

filter_df <- df %>%
  filter(windfg >= 0)

individuals_with_win <- n_distinct(winners$pidp)
total_individuals <- n_distinct(filter_df$pidp)
percentage_with_win <- (individuals_with_win / total_individuals) * 100
```

Summary Statistics		
Description	Paper	Replication
% of working-age individuals who report at least one lottery win	36%	25.34
% of small wins	81%	79.1%
% of medium wins	14%	15.2%
% of large wins	5%	5.76%
75th percentile win (top 25% win cutoff)	£85.66	£87.6
Average top 25% win	£831.16	£862
Average bottom 75% win	£25.69	£26.4
% self-employed	8.1%	8.18%
% self-employed that report at least one lottery win	36%	17.563%
Average inflation-adjusted win given self-employed	£595.50	£617
Average inflation-adjusted win given not self-employed	£192.79	£201
Average inflation-adjusted win for women given self-employed	£197.57	£209
Average inflation-adjusted win for women given not self-employed	£168.91	£176
Average inflation-adjusted win for men given self-employed	£696.92	£722
Average inflation-adjusted win for men given not self-employed	£211.99	£221

Figure 1: This table provides a comparison between the summary statistics from Fleche et al. and my own replication, along with a brief description on the left-hand side.

Similarly, Fleche et al. report that the proportion of individuals who report at least one win does not change when comparing self-employed people to the general population. They report that of self-employed people in the data, 36% of them report at least one lottery win. However, during my replication, I found only 17.5% of self-employed individuals report at least one lottery win.

```
SE_with_win <- selfemployed %>%
  group_by(pidp) %>%
  summarize(any_win = any(windfg == 1)) %>%
  ungroup()

t_individual <- n_distinct(selfemployed$pidp)
individuals_with_win_count <- SE_with_win %>%
  filter(any_win) %>%
  nrow()

percentage_with_win <- (individuals_with_win_count / t_individual) * 100
```

5 Figures and Tables

All regressions and figures (aside from figure 2) control for age, age squared, log real household income, marital status, health status, education, home ownership, and the number of days spent in the hospital in the previous year. Table 3 (Figure 6) additionally controls for region and wave.

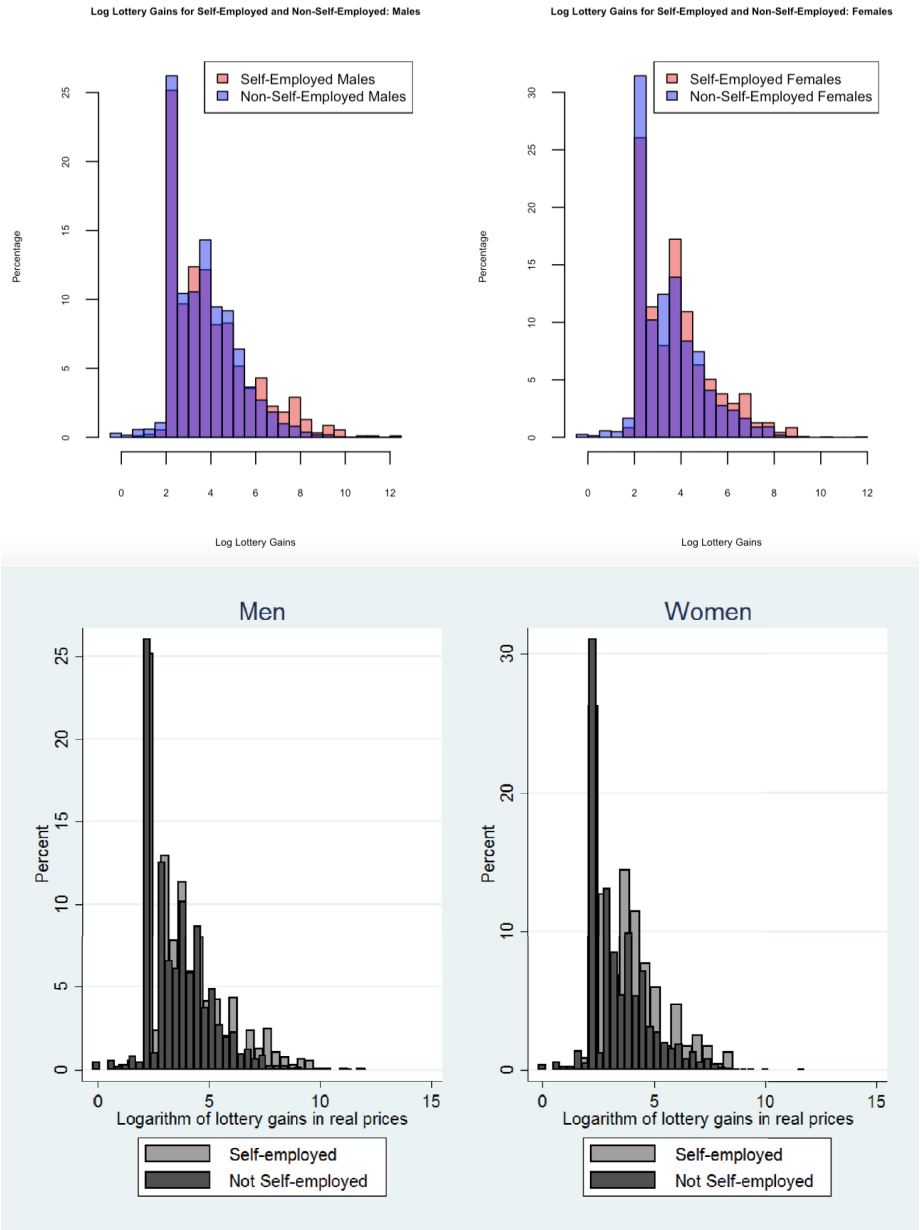


Figure 2: Replication (top) and Fleche et al. figure (bottom). Both graphs depict inflation-adjusted log lottery gains. The replication graph is a histogram with 30 breaks.

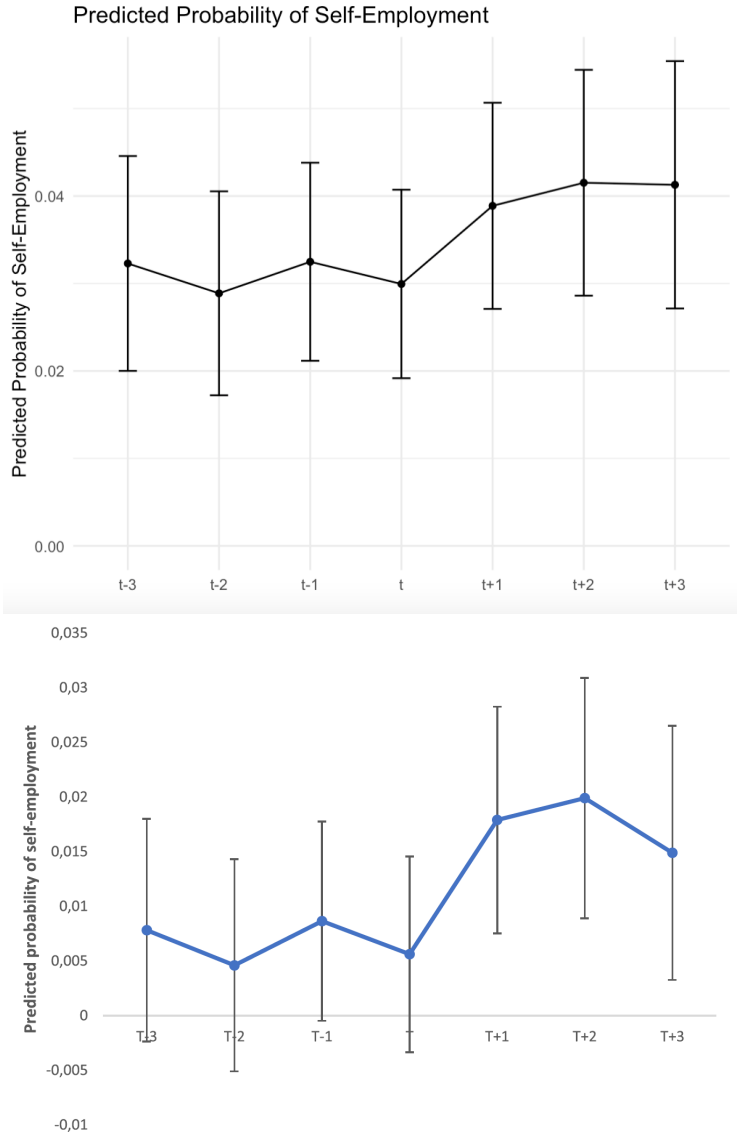
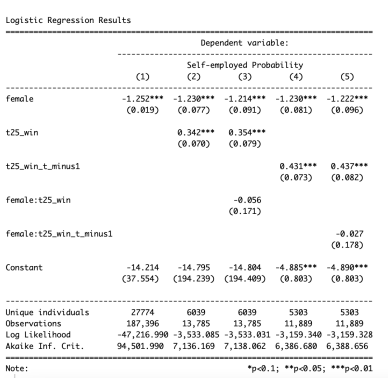


Figure 3: Replication (top) and Fleche et al. figure (bottom). Number of observations used in the estimation (replication): t 3 (N = 9,252), t 2 (N = 10,614), t 1 (N = 12,144), t (N = 13,904), t + 1 (N = 12,808), t + 2 (N = 11,841), and t + 3 (N = 10,990). Number of observations used in the estimation (Fleche et al.): t 3 (N = 10,818), t 2 (N = 11,661), t 1 (N = 12,629), t (N = 13,934), t + 1 (N = 11,952), t + 2 (N = 10,401), and t + 3 (N = 9013). While the overall shape of the two figures follows the same pattern, the replication coefficients seem to be about 0.02 higher than the coefficients in Fleche et al.. Additionally, while Fleche et al. show that lottery wins in period t do not have a statistically significant difference in the probability of self-employment in preceding periods, the replication graph has statistically significant coefficients in all periods.



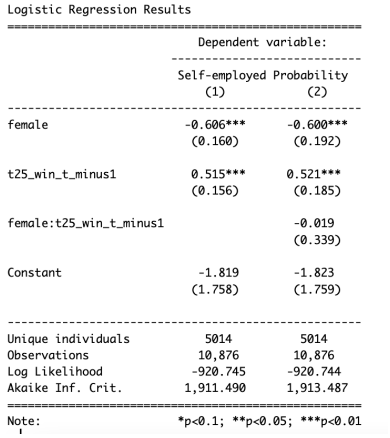
Logistic Regression Results

	Dependent variable:				
	Self-employed Probability				
	(1)	(2)	(3)	(4)	(5)
female	-1.252***	-1.230***	-1.214***	-1.230***	-1.222***
	(0.019)	(0.077)	(0.091)	(0.081)	(0.096)
t25_win		0.342***	0.354***		
		(0.070)	(0.079)		
t25_win_t_minus1				0.431***	0.437***
				(0.073)	(0.082)
female:t25_win			-0.056		
			(0.171)		
female:t25_win_t_minus1					-0.027
					(0.178)
Constant	-14.214	-14.795	-14.804	-4.885***	-4.890***
	(37.554)	(194.239)	(194.409)	(0.803)	(0.803)
Unique individuals	27774	6039	6039	5303	5303
Observations	187,396	13,785	13,785	11,889	11,889
Log Likelihood	-47,216.990	-3,533.085	-3,533.031	-3,159.340	-3,159.328
Akaike Inf. Crit.	94,501.990	7,136.169	7,138.062	6,386.680	6,388.656
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 1
Self-employment and lottery wins: Logit with random effects regressions.

	Self-employment probability in t					
	(1)	(2)	(3)	(4)	(5)	(6)
Female	−2.515***	−2.513***	−2.518***	−2.586***	−2.353***	−2.516***
	(0.0778)	(0.0778)	(0.271)	(0.291)	(0.386)	(0.409)
Top 25% winner in period t		0.230**	0.182	0.110		
		(0.107)	(0.155)	(0.169)		
Female × Top 25% winner in period t				0.259		
				(0.389)		
Top 25% winner in period $t-1$					0.661***	0.547***
					(0.180)	(0.193)
Female × Top 25% winner in period $t-1$						0.423
						(0.452)
Marginal effects at the mean						
Female	−0.072***	−0.072***	−0.061***	−0.061***	−0.060***	−0.062***
	(0.002)	(0.002)	(0.006)	(0.006)	(0.007)	(0.006)
Top 25% winner in period t		0.006**	0.005			
		(0.003)	(0.004)			
Male: Top 25% winner in period t				0.004		
				(0.006)		
Female: Top 25% winner in period t				0.004		
				(0.004)		
Top 25% winner in period $t-1$					0.018***	
					(0.005)	
Male: Top 25% winner in period $t-1$						0.022***
						(0.008)
Female: Top 25% winner in period $t-1$						0.014**
						(0.007)
Control variables	Yes	Yes	Yes	Yes	Yes	Yes
Regional dummies	Yes	Yes	Yes	Yes	Yes	Yes
Wave dummies	Yes	Yes	Yes	Yes	Yes	Yes
Observations	189,820	189,820	13,934	13,934	11,952	11,952
Number of individuals	28,042	28,042	6097	6097	5331	5331

Figure 4: Replicated table 1 (left) and table 1 of Fleche et al. (right). The control variables are hidden in both graphs. They are: age, the age squared, the log of real equivalent household income, dummies for marital status, dummies for self-reported health status, education dummies, home ownership, the number of days spent in hospital in year t-1 and the number of dependent children. Marginal Effects at the mean for the female variable are -0.0875, -0.0879, -0.0880, -0.0916, and -0.0916 for models 1 through 5, respectively. The marginal effect at the mean for t25 win is 0.0245 and 0.0244 in models 2 and 3, respectively. The marginal effect at the mean for t25 win in period t-1 is 0.0321 and 0.0320 in models 4 and 5, respectively.



Logistic Regression Results

	Dependent variable:	
	Self-employed Probability	
	(1)	(2)
female	-0.606***	-0.600***
	(0.160)	(0.192)
t25_win_t_minus1	0.515***	0.521***
	(0.156)	(0.185)
female:t25_win_t_minus1		-0.019
		(0.339)
Constant	-1.819	-1.823
	(1.758)	(1.759)
Unique individuals	5014	5014
Observations	10,876	10,876
Log Likelihood	-920.745	-920.744
Akaike Inf. Crit.	1,911.490	1,913.487
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2
Transition into self-employment and lottery wins in $t-1$: Logit with random effects regressions.

	Self-employment probability in t, conditioning on not being in self-employment in $t-1$	
	(1)	(2)
Female	−0.755***	−0.744***
	(0.224)	(0.256)
Top 25% winner in period $t-1$	−0.677***	−0.688***
	(0.194)	(0.234)
Female × Top 25% winner in period $t-1$		−0.0361
		(0.411)
Marginal effects at the mean		
Female	−0.010***	−0.010***
	(0.003)	(0.003)
Top 25% winner in period $t-1$	0.009***	
	(0.003)	
Male: Top 25% winner in period $t-1$		0.013***
		(0.005)
Female: Top 25% winner in period $t-1$		0.007*
		(0.004)
Control variables	Yes	Yes
Regional dummies	Yes	Yes
Wave dummies	Yes	Yes
Observations	10,620	10,620
Number of individuals	4890	4890

Figure 5: Replicated table 2 (left) and table 2 of Fleche et al. (right). The control variables are hidden in both graphs.

Logistic Regression Results

	Dependent variable:					
	Self-employed Probability					
	(1)	(2)	(3)	(4)	(5)	(6)
female	-0.498*	-0.742***	-0.824	-0.534**	-0.645**	-0.619**
	(0.281)	(0.269)	(0.588)	(0.207)	(0.275)	(0.276)
tZ5_win_t_minus1	0.504*	0.547**	0.539	0.610***	0.256	0.796***
	(0.270)	(0.258)	(0.502)	(0.202)	(0.264)	(0.265)
female:tZ5_win_t_minus1	-0.543	0.319	0.287	-0.100	0.808*	-1.450**
	(0.571)	(0.441)	(1.038)	(0.362)	(0.437)	(0.675)
Constant	-21.815	-10.258*	-22.197	-0.675	-24.147	-1.363
	(1,150.522)	(6.029)	(2,562.985)	(1.859)	(1,026.723)	(2.288)
Observations	4,863	6,013	1,247	9,381	6,369	4,507
Log Likelihood	-403.838	-491.063	-111.538	-776.341	-490.302	-398.713
Akaike Inf. Crit.	879.675	1,054.126	295.075	1,624.681	1,052.605	869.427
Note:						*p<0.1; **p<0.05; ***p<0.01

Logistic Regression Results

	Dependent variable:					
	Self-employed Probability					
	(1)	(2)	(3)	(4)	(5)	(6)
female	-0.238	-0.226	-0.699	0.478	-3.451	-1.020
	(0.618)	(0.794)	(0.689)	(0.686)	(2.830)	(2.257)
log_windfgy_t_minus1	0.125	0.224**	0.102	0.262***	0.128	0.364**
	(0.091)	(0.092)	(0.087)	(0.094)	(0.434)	(0.156)
female:log_windfgy_t_minus1	-0.079	-0.176	-0.047	-0.256	0.610	-0.376
	(0.159)	(0.200)	(0.177)	(0.171)	(0.594)	(0.617)
Constant	-3.261	-17.894	-0.906	-35.603	-41.335	-14.810
	(3.129)	(1,045.742)	(2.380)	(3,218.718)	(13,702.530)	(2,753.110)
Observations	5,818	4,186	6,509	3,495	243	1,743
Log Likelihood	-442.695	-317.202	-449.099	-312.639	-23.182	-86.711
Akaike Inf. Crit.	947.390	704.403	968.198	697.277	116.364	245.422
Note:						*p<0.1; **p<0.05; ***p<0.01

Table 3
Self-employment and lottery wins across different subsamples: Logit with random effects regressions.

	Self-employment Probability in t					
	Young	*Old*	*High education*	*Low education*	*High income*	*Low income*
Female	−2.097***	−2.918***	−2.432***	−2.716***	−2.633**	−2.513***
	(0.503)	(0.493)	(0.526)	(0.476)	(1.101)	(0.359)
Top 25% winner in period $t-1$	0.635*	0.595***	0.561*	0.572**	0.211	0.727***
	(0.330)	(0.224)	(0.289)	(0.256)	(0.313)	(0.261)
Female × Top 25% winner in	0.058	0.633	0.650	0.231	0.353	0.246
period $t-1$	(0.678)	(0.545)	(0.607)	(0.629)	(0.845)	(0.529)
Marginal effects at the mean						
Female	−0.036***	−0.090***	−0.062***	−0.062***	−0.062***	−0.069***
	(0.006)	(0.0092)	(0.0082)	(0.0107)	(0.019)	(0.0088)
Male: Top 25% winner in	0.017*	0.022**	0.022*	0.023**	0.008	0.032***
period $t-1$	(0.010)	(0.009)	(0.011)	(0.010)	(0.012)	(0.012)
Female: Top 25% winner in	0.007	0.027**	0.021*	0.010	0.008	0.015*
period $t-1$	(0.007)	(0.013)	(0.012)	(0.008)	(0.011)	(0.008)
Observations	5736	6216	4681	7148	5978	5974
Number of individuals	2889	2786	2220	3114	2964	3284
	Self-employment Probability in t					
	Partnered	*Not partnered*	*No children*	*With children*	*Renters*	*Homeowners*
Female	−2.478***	−2.279***	−2.663***	−2.564***	−4.035***	−2.366***
	(0.335)	(0.714)	(0.725)	(0.599)	(1.401)	(0.363)
Log (real lottery win) in	0.517**	1.220***	0.801***	0.435	0.948*	0.534***
period $t-1$	(0.202)	(0.445)	(0.240)	(0.355)	(0.547)	(0.201)
Female× Log (real lottery	0.471	−0.560	−0.0136	1.145	−0.636	0.478
win) in period $t-1$	(0.434)	(1.218)	(0.643)	(0.697)	(1.435)	(0.454)
Marginal effects at the mean						
Female	−0.076***	−0.031***	−0.054***	−0.061***	−0.033**	−0.065***
	(0.007)	(0.008)	(0.008)	(0.009)	(0.012)	(0.007)
Male: Top 25% winner in	0.021**	0.030**	0.030***	0.014	0.021	0.022***
period $t-1$	(0.008)	(0.013)	(0.009)	(0.012)	(0.015)	(0.008)
Female: Top 25% winner in	0.0208**	0.006	0.008	0.016**	0.002	0.019**
period $t-1$	(0.009)	(0.011)	(0.006)	(0.008)	(0.010)	(0.009)
Observations	8750	3202	8143	3809	2565	9387
Number of individuals	3894	1716	3834	1884	1437	4142

Figure 6: Replicated table 3 (top) and table 3 of Fleche et al. (bottom). Models 1 through 6 (young, old, high education, low education, high income, and low income) are shown in the top left table while models 7 through 12 are shown in the top right table. The control variables are hidden in both graphs. From the footnote of Fleche et al.: The dependent variable is a dummy variable that takes the value one if the respondent is self-employed in period t and zero otherwise. "Young" and "Old" are, respectively, respondents below age 40 and between 40 and 65 years old. "High Education" and "Low Education" are, respectively, respondents with and without a university degree. "Low Income" and "High Income" are, respectively, respondents with a log of real equivalent household income below and above the median of the log of real equivalent household income of the estimation sample. The figures refer to the lottery winners a year after winning who were not self-employed in the previous period (period t minus 1). These models also control for wave and region.

When replicating Table 3, I restricted the data to those who reported being employed in the previous period by filtering out all data points that reported being self-employed in period t-1. I then created sub data sets that conform to the restrictions outlined in the footnotes of the graph and ran the same regression on the relevant data set. Here is the standardized code:

```
#For models 1 through 6; sample code from model 1 (Young):
youngdf <- subset(table2df, age < 40)
t3_m1 <- glm(selfemployed ~ female * t25_win_t_minus1 + age +
I(age^2) + log_fihhyr + mastat + nkids + hlstat + wave + region,
data = youngdf, family = binomial())

#For models 7 through 12; sample code from model 12 (Homeowners):
homeowners <- subset(table2df, hsowrp ==1)
t3_m12 <- glm(selfemployed ~ female * log_windfgy_t_minus1 + age +
I(age^2) + log_fihhyr + mastat + nkids + hlstat + wave + region,
data = homeowners, family = binomial())
```

Note that for models 7 and 9 (Partnered and No Children), I did not control for master and kids, respectively, since the variable only contained a single value due to the data restriction. All other models contain the same control variables as the models shown above. I would also like to note that Fleche et al. describe "High Education" as people who have received a university degree. The variable qfedhi takes the values: "Higher Degree", "First Degree", "Teaching QF", "Other Higher QF", "Nursing QF", "GCE A Levels", "GCE O Levels or Equiv", "Commercial QF, No O Levels", "CSE Grade 2-5, Scot Grade 4-5", "Apprenticeship", "Other QF", "No QF", "Still At School No QF", "don't know", "missing or wild", "inapplicable", "proxy", "refused". I restricted higher education to those who report either a Higher Degree or First Degree.

6 Conclusion

This replication of the study by Fleche et al. reveals both substantial similarities and notable differences compared to the original findings. In terms of overall trends, the replication confirms that lottery winnings—especially those in the top 25th percentile—are associated with a higher likelihood of self-employment. The summary statistics related to lottery sizes, such as the proportions of small, medium, and large lottery wins, align closely with the original study's findings.

However, there are key differences in the specific details of the analysis. Most notably, my replication found that only 25% of working-age individuals in the BHPS dataset reported at least one lottery win, compared to the 36% reported by Fleche et al. When narrowing the focus to the years when lottery data became available (post-1997), this percentage increased slightly to 26.23%, but still remained significantly lower than the original estimate. Similarly, among the self-employed population, only 17.5% reported at least one lottery win in my replication, compared to 36% in the original paper. These discrepancies suggest

that there may have been differences in how lottery wins and self-employment statuses were operationalized or filtered in the original analysis. Additionally, while Fleche et al. found no significant changes in the proportion of lottery winners when comparing self-employed individuals to the general population, my replication detected some deviations, particularly in earlier waves.

Several methodological issues in the original paper also merit attention. First, the use of data from all 18 waves despite lottery data only being available from wave 8 raises concerns about the consistency and accuracy of the comparisons drawn between waves. Including data from earlier waves without corresponding lottery data could introduce bias, as the study may inadvertently compare non-lottery winners from earlier waves with lottery winners from later waves. This issue complicates the analysis and weakens the conclusions about the impact of lottery winnings on self-employment over time. Second, the original study does not provide sufficient detail regarding the filtering and operationalization of key variables, such as self-employment status and lottery win reporting. This lack of transparency makes it difficult to assess the accuracy of the comparisons and may account for some of the differences in findings between the original study and this replication. For instance, my replication of the second figure in Fleche et al. (Figure 3) has a similar shape to the overall trend shown, but shows statistically significant differences in probability of being self employed across all periods (from t-3 to t+3), while the original figure only shows a statistically significant difference proceeding a lottery win, suggesting a potential methodological or data-handling discrepancy.