ECON 453
In-Class Exercise 7
October 24, 2023

ANSWER KEY

Please download the file "IC7.gdt", a gretl data file. This dataset comes from the 2019 American Community Survey and is the same data we used in Problem Set 2 and In-Class Exercise 6. The dataset includes individuals that have a bachelor's degree in economics, accounting, marketing, or finance, work at least 30 hours per week, make at least $15,000 per year, and are between the ages of 25 and 40. Please open the data file. The dataset contains basic descriptions of each of the variables.

In addition, please download and open the Excel file called "IC7 Prediction Template". You will find two sheets, one for the Linear Probability Model, and one for the Logit model.

We will follow what we did in In-Class Exercise 6, and use "**ever_married**" as our dependent variable. This variable is 0 if the individual has never been married and 1 if they have been married at some point in their life.

1. Run a linear probability model (OLS regression) using the "Ever married" variable as the dependent and the following regressors: female, age, immig, and nchild.
    a. Use the LPM prediction template in Excel. Plug in the coefficients from your regression.
    b. Last time, we were predicting for Janky McMurphy, a 33-year-old female Irish Immigrant. Predict the following probabilities

Here are my regression results:

Model 1

File Edit Tests Save Graphs Analysis LaTeX

```
Model 1: OLS, using observations 1-13114
Dependent variable: ever_married

             coefficient   std. error    t-ratio    p-value
  -----------------------------------------------------------
  const      -0.414724     0.0282235     -14.69    1.70e-048 ***
  female      0.0334998    0.00722952      4.634   3.63e-06  ***
  age         0.0264480    0.000904390    29.24    3.53e-182 ***
  immig       0.0626495    0.00997395      6.281   3.46e-010 ***
  nchild      0.179282     0.00387161     46.31    0.0000    ***

Mean dependent var   0.599817   S.D. dependent var   0.489954
Sum squared resid    2178.134   S.E. of regression   0.407622
R-squared            0.308054   Adjusted R-squared   0.307843
F(4, 13109)          1459.032   P-value(F)           0.000000
Log-likelihood      -6836.756   Akaike criterion     13683.51
Schwarz criterion    13720.92   Hannan-Quinn         13696.01
```

Then I used the prediction template file in Excel (see the template in Canvas):

C	D	E	F	G	H
Variable	**Coefficient**	**Value 1**	**Value 2**	**Value 3**	**Value 4**
const	-0.414724	--	--	--	--
female	0.0334998	1	1	1	1
age	0.026448	33	33	33	33
immig	0.0626495	1	1	1	1
nchild	0.179282	0	1	2	3
	Pred Probabilit	55.4%	73.3%	91.3%	109.2%
	Estimated Marginal Effects		17.93	17.93	17.93
			"Percentage Points"		

i. The probability Janky has been married if she has 0 children: **55.4%**

ii. The probability Janky has been married if she has 1 child: **73.3%**

iii. The probability Janky has been married if she has 2 children: **91.3%**

iv. The probability Janky has been married if she has 3 children: **109.2%**

c. Comment briefly on how realistic the estimated marginal effects of having children are in the linear probability model.

I think there are a couple of issues that should draw your attention on this one. The first is that the estimate with 3 children gives us an unreasonable prediction of over 100%. The other, more complicated, issue is that this model predicts linear effects from each additional child. We know this is a linear model, and that our estimated impact of each child is that the likelihood of having been married increases by about 18 percentage points. To me, that is not very realistic. I think that the big difference is between having children or not. In other words, if you told me a person has 3 children and a second person with identical attributes has 2 children, I would estimate about the same probability that these two people are (or have been) married. This is not what the model predicts.

2. Run the same regression as in question 1 but use a binary Logit model instead.
   a. Use the Logit prediction template in Excel. Plug in the coefficients from your regression.
   b. Find the mean values for each of the explanatory values and plug them into the “Means” column in the template.
   c. Find the “marginal effects” of being female and being an immigrant. To do this, use the mean value for each variable, then adjust the dummy variable of interest from 0 to 1.

My regression results from the binary Logit model:

```
Model 2

File  Edit  Tests  Save  Graphs  Analysis  LaTeX

Model 2: Logit, using observations 1-13114
Dependent variable: ever_married
Standard errors based on Hessian

             coefficient   std. error      z        slope
  ---------------------------------------------------------
  const       -4.41368     0.171021     -25.81
  female       0.109193    0.0448706      2.433   0.0211202
  age          0.124458    0.00545780    22.80    0.0241750
  immig        0.359028    0.0611157      5.875   0.0655162
  nchild       1.74451     0.0464120     37.59    0.338858

Mean dependent var   0.599817   S.D. dependent var   0.489954
McFadden R-squared   0.306003   Adjusted R-squared   0.305436
Log-likelihood      -6125.810   Akaike criterion     12261.62
Schwarz criterion    12299.03   Hannan-Quinn         12274.11
```

Again, I used the template to find the marginal effects. For "female" you put all the other variables at their average, then change the female from 0 to 1. For "immig", do the same thing. You need to do these one at a time to isolate the effects.

For female:

Variable	Coefficient	mean	Value 1	Value 2
const	-4.41368		--	--
female	0.109193	0.4197	0	1
age	0.124458	32.38	32.38	32.38
immig	0.359028	0.1522	0.1522	0.1522
nchild	1.74451	0.7504	0.7504	0.7504
		Exp Function	2.6644413	2.971858
		Pred Probabilit	72.7%	74.8%
	Estimated Marginal Effects			2.11

For immig:

Variable	Coefficient	mean	Value 1	Value 2
const	-4.41368		--	--
female	0.109193	0.4197	0.4197	0.4197
age	0.124458	32.38	32.38	32.38
immig	0.359028	0.1522	0	1
nchild	1.74451	0.7504	0.7504	0.7504
		Exp Function	2.6410555	3.781825
		Pred Probabilit	72.5%	79.1%
	Estimated Marginal Effects			6.55

i. Marginal effect of being female: **2.11 percentage points**

ii. Marginal effect of being an immigrant: **6.55 percentage points**

OMG, these are the "slopes" that are displayed in the gretl results!!!!!!!

d. Last time, we were predicting for Janky McMurphy, a 33-year-old female Irish Immigrant. Predict the following probabilities

Let me use my patented prediction template:

Variable	Coefficient	mean	Value 1	Value 2	Value 3	Value 4
const	-4.41368		--	--	--	--
female	0.109193	0.4197	1	1	1	1
age	0.124458	32.38	33	33	33	33
immig	0.359028	0.1522	1	1	1	1
nchild	1.74451	0.7504	0	1	2	3
		Exp Function	1.1754546	6.72724	38.50065	220.3429
		Pred Probabilit	54.0%	87.1%	97.5%	99.5%
	Estimated Marginal Effects			33.03	10.41	2.08

i. The probability Janky has been married if she has 0 children: **54.0%**

ii. The probability Janky has been married if she has 1 child: **87.1%**

iii. The probability Janky has been married if she has 2 children: **97.5%**

iv. The probability Janky has been married if she has 3 children: **99.5%**

e. Compare the estimated marginal effects of having children in the Logit model to those from the linear probability model.

The Logit is a non-linear model, so it allows for the possibility that the effect of increasing a variable depends on the level of the variable. We see that the first child increases the likelihood of having been married by about 33 percentage points. The second child only increases the probability by 10.4% points, and the third one by about 2 % points. This matches more closely with what I would expect. The other thing we notice is that the Logit model does not allow for "illegal" probability predictions.

3. Create dummy variables for people that have 1 child in the home, 2 children in the home, and 3 or more children in the home. Run a linear probability model using "**ever_married**" as the dependent variable and the following regressors: female, age, immigrant, and your new children dummies.
   a. What is the estimated impact on the probability of being married from having the:

Here are the results from my regression:

Model 3

File Edit Tests Save Graphs Analysis LaTeX

```
Model 3: OLS, using observations 1-13114
Dependent variable: ever_married

                   coefficient   std. error    t-ratio    p-value
  -----------------------------------------------------------------
  const            -0.316569     0.0274795     -11.52     1.46e-030 ***
  female            0.0153160    0.00699261      2.190    0.0285    **
  age               0.0220257    0.000886013    24.86     2.45e-133 ***
  immig             0.0560833    0.00961938      5.830    5.67e-09  ***
  one_kid           0.447438     0.0102337      43.72     0.0000    ***
  two_kids          0.472214     0.0103480      45.63     0.0000    ***
  threeplus_kids    0.478207     0.0144983      32.98     2.85e-229 ***

Mean dependent var   0.599817   S.D. dependent var   0.489954
Sum squared resid    2024.565   S.E. of regression   0.393020
R-squared            0.356840   Adjusted R-squared   0.356546
F(6, 13107)          1212.011   P-value(F)           0.000000
Log-likelihood      -6357.348   Akaike criterion     12728.70
Schwarz criterion    12781.07   Hannan-Quinn         12746.19
```

      i. First child: **44.74 percentage points**

      ii. Second child: **2.48 percentage points**

      iii. Third child: **0.6 percentage points**

To find these, you use the coefficients and remember that each is compared to the reference category (no children). The "one_kid" coefficient is a direct estimate of the effect of the first child. To find the effect of the second child (specifically), we need to subtract the impact of having two kids minus the impact of having one kid (0.4722 – 0.4474). As with the Logit model, this version of the model predicts that the question of whether a person has kids is much more important than how many kids in terms of predicting likelihood of marriage.

4. Create a simple dummy variable, “kids”, that is 1 if the person has any children in the home and 0 if not. Create an interaction term between the female and “kids” variables. Run a linear probability model using “**ever_married**” as the dependent variable and the following regressors: female, age, immigrant, kids, and the interaction term.
    a. Report the coefficient on the interaction term. What does this tell us, and does this make sense?

Model 4

File Edit Tests Save Graphs Analysis LaTeX

```
Model 4: OLS, using observations 1-13114
Dependent variable: ever_married

               coefficient   std. error    t-ratio    p-value
  -------------------------------------------------------------
  const        -0.336428     0.0270825     -12.42     3.12e-035 ***
  female        0.0550586    0.00910629      6.046    1.52e-09  ***
  age           0.0221631    0.000871403    25.43     2.48e-139 ***
  immig         0.0563328    0.00959111      5.873    4.37e-09  ***
  kids          0.505329     0.0101831      49.62     0.0000    ***
  female_kids  -0.0974647    0.0141321      -6.897    5.57e-012 ***

Mean dependent var   0.599817   S.D. dependent var   0.489954
Sum squared resid    2018.114   S.E. of regression   0.392378
R-squared            0.358889   Adjusted R-squared   0.358645
F(5, 13108)          1467.552   P-value(F)           0.000000
Log-likelihood      -6336.424   Akaike criterion     12684.85
Schwarz criterion    12729.74   Hannan-Quinn         12699.84
```

The coefficient on the female/kids interaction term is negative and significant. This tells us that the presence of children in the household had less of an effect on the likelihood a female is married as compared to the likelihood a male is married. Another way of thinking about it: the impact of kids on likelihood of marriage for males is about 50.5 percentage points, for females the estimated impact is about 40.8 percentage points.

5. Run the same regression as in question 4 but use a binary Logit model instead.
    a. Compare the estimated interaction effect from this model to the one in the LPM in question 4.

Model 5

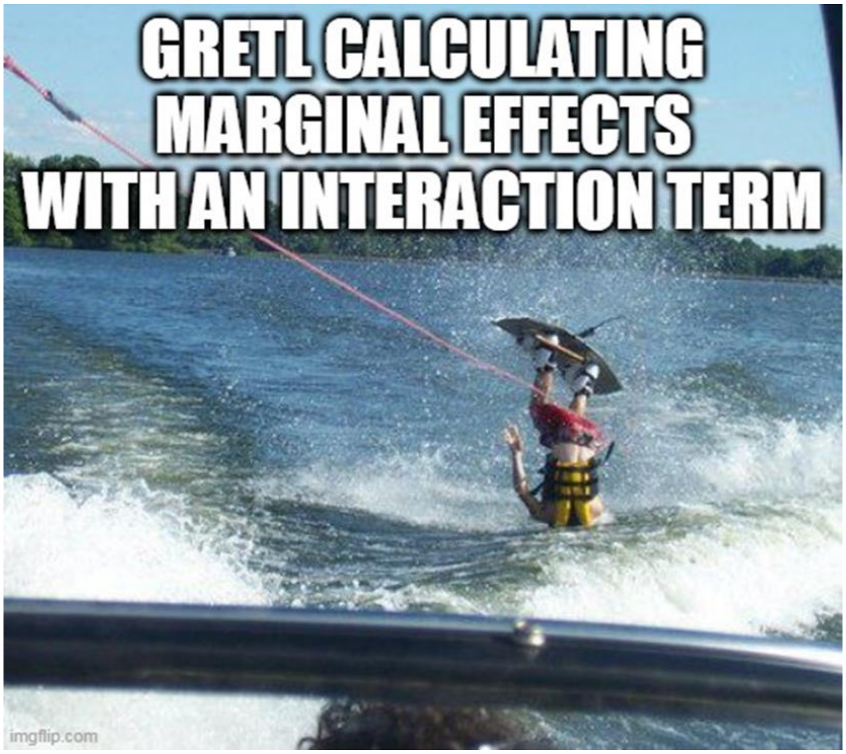
File Edit Tests Save Graphs Analysis LaTeX

```
Model 5: Logit, using observations 1-13114
Dependent variable: ever_married
Standard errors based on Hessian

               coefficient   std. error      z         slope
  -------------------------------------------------------------
  const        -4.63230      0.173234     -26.74
  female        0.243588     0.0496303      4.908     0.0512021
  age           0.128329     0.00550594    23.31      0.0272010
  immig         0.351796     0.0621184      5.663     0.0707948
  kids          3.29375      0.0998466     32.99      0.567082
  female_kids  -1.02962      0.127612      -8.068    -0.237635

Mean dependent var    0.599817   S.D. dependent var   0.489954
McFadden R-squared    0.314376   Adjusted R-squared   0.313696
Log-likelihood       -6051.902   Akaike criterion     12115.80
Schwarz criterion     12160.69   Hannan-Quinn         12130.80
```

From the “slope” column we should notice that the interaction effect is much larger in the Logit model than the linear probability model. This is one of the reasons why we might use the “simpler” LPM. Certain modeling techniques we are used to, such as interacting variables, do not work perfectly in non-linear estimation techniques like a Logit. Allow me to summarize with a meme:


GRETL CALCULATING MARGINAL EFFECTS WITH AN INTERACTION TERM
imgflip.com