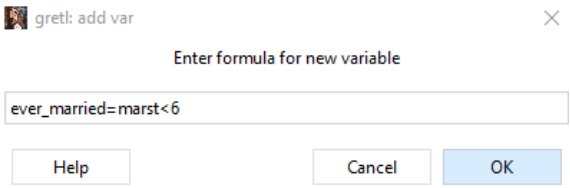
ECON 453
In-Class Exercise 6
October 19, 2023

ANSWER KEY

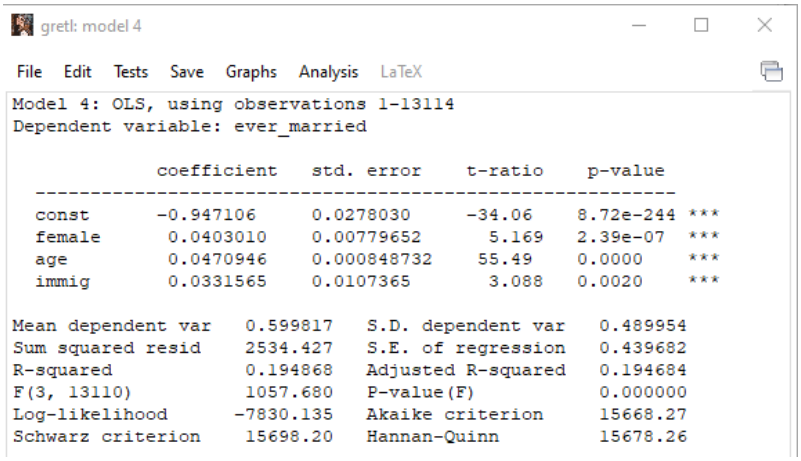
Please download the file "IC6.gdt", a gretl data file. This dataset comes from the 2019 American Community Survey and is the same dataset we used in Problem Set 2. The data contain information at the individual level on work outcomes and demographic information. The dataset we are looking at today includes 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.

1. We are going to work with a binary dependent variable, whether or not the individual has ever been married in their (short) life. Please define this variable using the "marst" variable in the data.
    a. Run a linear probability model (OLS regression) using the "ever married" variable as the dependent and the following regressors: female, age, immig.
        i. Interpret the coefficients and discuss briefly whether these make sense.

The first step is to create the binary "ever married" variable. My command to do this in gretl was:



Once I did that, here are the results of my linear probability model:

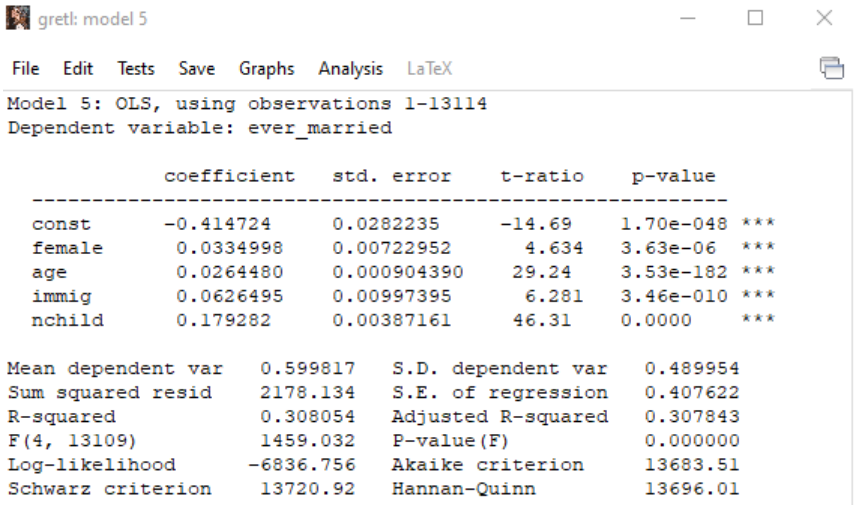


Since we are predicting a probability, we want to be using the term <u>percentage points</u> in our interpretations. We can interpret the coefficients as:

- ***Females (25 to 40 with business degree) are about 4 percentage points more likely to have been married than males***
- ***Each year of age increases likelihood of being married by about 4.7% points (be careful out there guys)***
- ***Those born outside of the U.S. are about 3.3 percentage points more likely to have been married (or still be married)***

In terms of whether these make sense, it is a little odd to me that the number is so much different for males and females. I think this mostly has to do with the fact that we are looking at a specific age range (25 to 40 years old) and a specific type of person (has a bachelor's degree in accounting, economics, finance, or marketing). Females are more likely to be married younger, so I am guessing this explains a lot of the difference. The fact that marriage probability increases with age doesn't surprise me.

b. Run the model again and add the "nchild" variable as an additional regressor.



gretl: model 5

File Edit Tests Save Graphs Analysis LaTeX

```
Model 5: 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
```

i. Interpret the coefficient on the nchild variable

Every child that a person has living with them increases the probability they are (or have been) married by about 18 percentage points.

ii. Predict the probability that Janky McMurphy is/was married. She is a 33-year-old female that immigrated from Ireland. Janky has 4 kids and a heart of gold. Comment on your prediction.

I would like to hear more about Janky and her life choices, but for now, let's just predict the probability she has been married at some point in her (interesting!) life:

$$\hat{y} = -0.4147 + (0.0335 * 1) + (0.0264 * 33) + (0.0626 * 1) + (0.1793 * 4) = 1.27$$

Our model predicts there is a 127% chance that Janky has been married (or is currently married). You should be alarmed by predicting a probability over 100%.

iii. Was the addition of nchild an improvement to our model?

There are a couple of issues to consider for this one. The first thing we might look at is the statistical significance of the "nchild" variable. This variable is highly statistically significant, which tells us it is useful in our model. The second thing we might look at is the explanatory power of our models. Our adjusted R^2 without the nchild variable was around 0.194. When we added it, that statistic jumped to 0.304. This is another indication that this was a good idea.

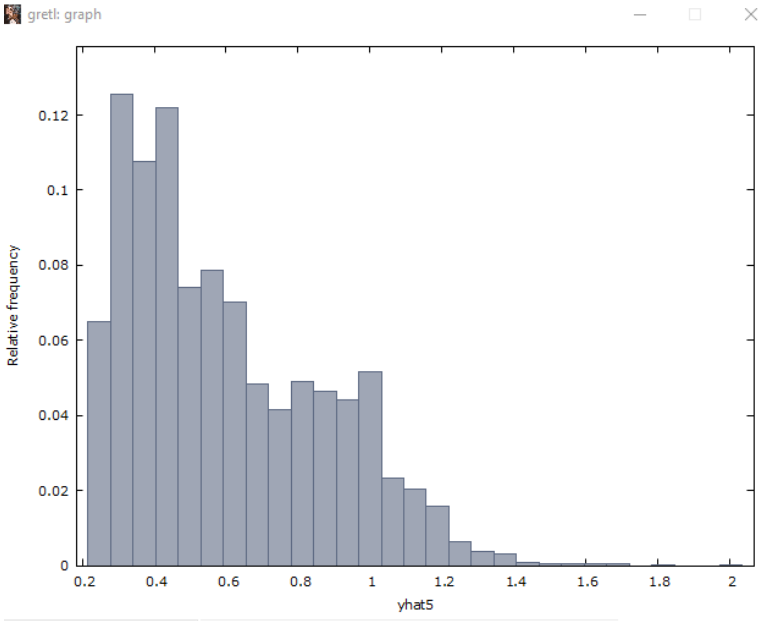
The other issue to think about, however, is whether this is a good idea in terms of how we are meant to specify models. When we estimate regressions, there is an implication that causality runs from the x-variable to the y-variable. In this case, that would mean that having more children makes one more likely to get/be married. In my mind, it makes more sense that those that are (or have been) married are more likely to have children. Overall, I am torn about adding this variable to the model.

c. Save the fitted values from your regression in part b (from the results menu, select **Save -> Fitted values**). Generate the summary statistics and frequency distribution for these fitted values.
    i. Comment on what these tell us about our estimated probabilities.

gretl: summary stats: yhat5

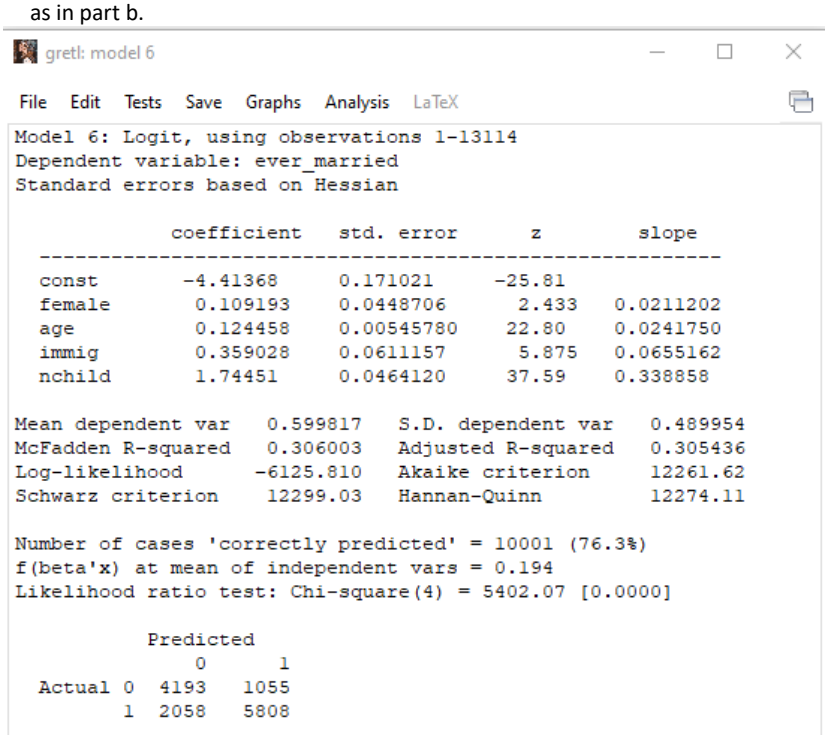
```
Summary statistics, using the observations 1 - 13114
for the variable 'yhat5' (13114 valid observations)

  Mean                    0.59982
  Median                  0.53740
  Minimum                 0.24648
  Maximum                  2.0052
  Standard deviation      0.27194
  C.V.                    0.45337
  Skewness                0.73467
  Ex. kurtosis           -0.23727
  5% percentile           0.27292
  95% percentile           1.0979
  Interquartile range     0.42437
  Missing obs.                  0
```



I don't know about you guys, but these are the kinds of things that keep me awake at night. We should be alarmed at what these things are telling us. From the summary statistics, we can see that, for one of our observations, we predict an over 200% chance the person is married. The 95th percentile is over 1, so at least 5% of our observations have predictions of greater than 100%. The histogram also shows that there are a significant number of observations for which we have made unrealistic predictions. This is a sign that we either need to adjust our model specification (change variables, restrict sample, etc.) or we need to implement something like a Logit or Probit model instead.

d. Now run the same regression using a binary Logit model (instead of OLS). Use the same regressors as in part b.



gretl: model 6

File Edit Tests Save Graphs Analysis LaTeX

```
Model 6: 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

Number of cases 'correctly predicted' = 10001 (76.3%)
f(beta'x) at mean of independent vars = 0.194
Likelihood ratio test: Chi-square(4) = 5402.07 [0.0000]

           Predicted
               0      1
  Actual 0  4193   1055
         1  2058   5808
```

i. Compare the predicted impacts of each variable to what we saw in part b.

When we interpret the results of Logit models, we want to look at the "slopes at means" calculations, rather than the values in the "coefficient" column. Allow me to create a table to help us compare:

	Estimated % point Difference	
	LPM	**Logit**
Female	0.033	0.021
Age	0.026	0.024
Immigrant	0.063	0.066
# of Chilidren	0.179	0.339

The estimated impacts of the age and immigrant variables are strikingly similar across models, and the impact of being female is fairly close. The big difference in this case is the estimated impact from the number of children. The fact that these are so different comes from a couple of factors. First, this is a variable with a large impact. Second, the Logit "marginal effects" are calculated as the slope at the mean values of each variable. The mean value of kids in the sample is less than 1, so the Logit effect is telling us there is a tremendous change in the probability of marriage from the first child. This makes sense. The difference is probably between having children and not. In my mind, I don't think there is much difference in the impact of the second and third child on the likelihood of being married.

ii. Summarize what we learn from the "% correctly predicted" table in the results. What percentage of the sample has been married? What percentage of the sample does our model predict has been married?

We learn that the model estimates "correctly" for 76.3% of the sample. The table tells us that there are 7,866 people that are married (2058 + 5808). This is out of 13,114 total in the sample, so right about 60.0% of the sample is (or has been) married. Our model predicts that 6,863 people are married (1055+5808, from the table). The model correctly predicts for 10,001 people (4193 correctly predicted to never have been married, 5808 correctly predicted to have been married).

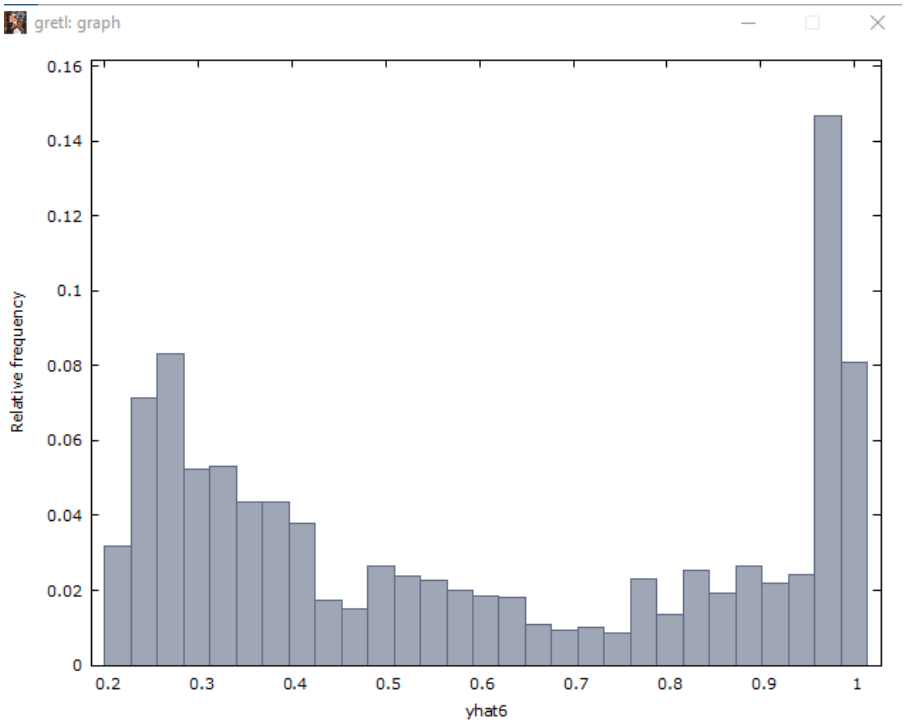
e. Save the fitted values from your regression in part d. Generate the summary statistics/frequency distribution for these fitted values.

i. Comment on how these compare to what we saw in part c.

gretl: summary stats: yhat6

Summary statistics, using the observations 1 - 13114
for the variable 'yhat6' (13114 valid observations)

Mean	0.59982
Median	0.51669
Minimum	0.21379
Maximum	1.0000
Standard deviation	0.28821
C.V.	0.48050
Skewness	0.18174
Ex. kurtosis	-1.6002
5% percentile	0.23271
95% percentile	0.99555
Interquartile range	0.60155
Missing obs.	0



These demonstrate one of the advantages of using the Logit model over the LPM. This type of model makes it impossible to get impossible predicted values (cannot have predicted probabilities of <0 or >1). You will notice in the histogram that there is a large bunching right around a probability of 1 – these are the people in the sample with a large number of children (people like Janky).

f. Using your model in part d, what is your predicted probability that Janky McMurphy has been married

Fun! Here we need to use the Logit functional form and plug in the values for Janky. Remember that we are using the "coefficients" in this equation, not the "slopes" that we use for interpretations.

$$\hat{y} = \frac{e^{-4.414+0.109+(0.124*33)+0.359+(1.745*4)}}{1+e^{-4.414+0.109+(0.124*33)+0.359+(1.745*4)}} = \frac{1243.9}{1244.9} = 0.999 = 99.9\%\ chance$$

Our Logit model predicts there is a 99.9% chance that Janky is married. I think this is probably factoring in her "heart of gold".