Andrew Plum

**ECON 453**

Fall 2023

Problem Set 3 – 38 points

Submit by end of day Monday November 6th

Please download the gretl data file “PS3 Data” from Canvas. This is a dataset from the 2021 General Social Survey (GSS). This is a survey of around 4,000 individuals that has been conducted on a regular basis since 1972. The data contain information at the individual level regarding demographic information as well as attitudes towards social issues, religious beliefs, and various other information. The data file has a brief description/list of codes for each variable, but please let me know if you have any questions as you work with the data.

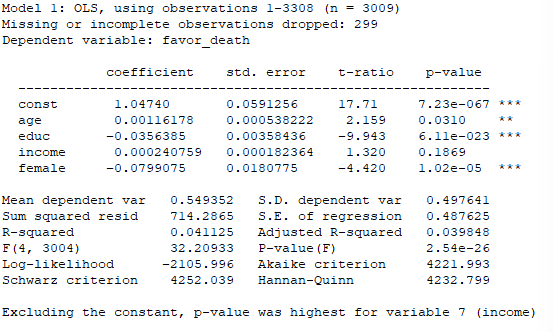
You will submit one document (Word or PDF) for this problem set. Please copy/paste the relevant regression results or graphs into your document, then add your discussions.

1. (7 points) To begin, we will focus on a binary dependent variable, whether the individual favors or opposes the death penalty as a form of punishment. The variable **favor\_death** will be the dependent variable for the first set of questions.

* Run a linear probability model using **favor\_death** as the dependent variable and the following regressors: female, age, educ (years of educ), and income.

Done

* Report/copy your results. Summarize what we learn from the model. Interpret the coefficients on the female and education variables, specifically. Overall, do the estimated coefficients match your expectations? Explain briefly.

 From the model, we learn several things. The regressor education is quite statistically significant and each additional year of education decreases the likelihood of a person favoring the death penalty by about 3.56 percentage points. The regressor females is also very statistically significant with women being roughly 8 percentage points less likely than men to support the death penalty, everything else equal. Age is still significant but less significant than education and females, and it suggests a small increase in the likelihood of favoring the death penalty as people grow older. The regressor income has a small positive coefficient and is not statistically significant. I am not surprised by the coefficients for education and females. In my personal experience, I have heard more arguments against the death penalty from women than men. I also know a lot of intellectuals argue against the death penalty. Although its tiny, I think the small increase in a person’s support for the death penalty might be explained by people becoming protective of their loved ones as they age, and this is because young adults which are in the sample likely haven’t started their own family yet but once they do, I would imagine their views in life change and this might include views on the death penalty. The lack of significance the regressor income has doesn't surprise me; I can’t see how it would factor into a person's views on the death penalty.

* Predict the probability that a 75-year-old male with an income of $60,000 per year is in favor of the death penalty. Compute this for 5 different levels of education: 10, 12, 14, 16, and 18 years. Report the probabilities and comment briefly if this seems like a reasonable set of marginal effects of education.

Probability that a 75-year-old male with an income of $60,000 per year with variable years of education supports the death penalty:

For 10 years of education: probability = 1.04740 + 0.00116178 \* 75 + (-0.0356385) \* 10 + 0.000240759 \* 60 - 0.0799075 \* 0 = 1.04740 + 0.0871335 - 0.356385 + 0.01444554 - 0 = 0.79259404

For 12 years of education: probability = 1.04740 + 0.00116178 \* 75 + (-0.0356385) \* 12 + 0.000240759 \* 60 - 0.0799075 \* 0 = 1.04740 + 0.0871335 - 0.427662 + 0.01444554 - 0 = 0.72131704

For 14 years of education: probability = 1.04740 + 0.00116178 \* 75 + (-0.0356385) \* 14 + 0.000240759 \* 60 - 0.0799075 \* 0 = 1.04740 + 0.0871335 - 0.498939 + 0.01444554 - 0 = 0.65004004

For 16 years of education: probability = 1.04740 + 0.00116178 \* 75 + (-0.0356385) \* 16 + 0.000240759 \* 60 - 0.0799075 \* 0 = 1.04740 + 0.0871335 - 0.570216 + 0.01444554 - 0 = 0.57876304

For 18 years of education: probability = 1.04740 + 0.00116178 \* 75 + (-0.0356385) \* 18 + 0.000240759 \* 60 - 0.0799075 \* 0 = 1.04740 + 0.0871335 - 0.641493 + 0.01444554 - 0 = 0.50748604

The probabilities for this man with 10, 12, 14, 16, 18 years of education are 0.79259404, 0.72131704, 0.65004004, 0.57876304, and 0.50748604 respectively. This pattern of probabilities aligns with what I’d expect, as higher education levels are generally associated with more liberal views on social issues. The marginal effects of education for when it changes by 2 years is always -0.071277. However, I don’t know if the impact of 2 years more of education should always be the same. For example, I would think the change between dropping out of high school and finishing high school and the change between finishing a 2 year degree and finishing a 4 year degree would be different and not the same as it is here.

1. (4 points) Let’s focus on the predicted values from the model in question 1.

* Create a histogram (frequency distribution) of the predicted values for all of the individuals in the sample. Comment on what we learn from this.

A graph of a graph

Description automatically generated

We learn from the histogram that most of the predicted values fall between the range of 30% to a bit under 80%. We also see that some of the predicted probabilities don’t make sense as some are above 100%.

* Create a table that shows the “percent correctly predicted” for the regression in question 1. To do this, you should use a benchmark value of 0.5. In the end, you want to show a cross-tabulation (**View -> Cross Tabulation**) between the actual and predicted values for the **favor\_death** variable.

A white paper with black text

Description automatically generated

* What percentage of the sample does our model correctly predict for? Which type of mistake is more common in our predictions?

From the cross tabulation, we can see that our model correctly predicted that 1219 people favor the death penalty and that 568 people did not. The model correctly predicted about 59.39% of the sample (calculated by (568 + 1219) / 3009). It was more common for the model to predict that a person favors the death penalty when they really do not (788 people from the sample) than for the model to predict that a person does not favor the death penalty when they really do (434 people from the sample).

1. (7 points) For this question, run a binary Logit model that uses the same set of variables as you used in Question 1.

* Report/copy your results.

A screenshot of a computer

Description automatically generated

* Compare the estimated effects of the variables from this model to those from the linear probability model used in question 1. Are there major differences in the models?

The values of the slope from the logit regression model are quite similar to their respective coefficients in the linear probability model. For example, the value of the age slope in the logit model was 0.00119300 and the respective coefficient in the linear probability model was 0.0016178 which is a difference of 0.00003122 which is small when considering the values both models gave for age for its estimated effects. The small difference in values between the logit and LPM were the case for the other variables as well. There were really no big differences between the regression results when it came to things like significance, sign, magnitude, etc.

* Predict the probabilities for the 75-year-old male with income of $60,000 for 5 education levels, as you did in question 1. Compare the marginal effects of increasing education in this model and discuss which seems more reasonable.

A screenshot of a spreadsheet

Description automatically generated

The predicted probabilities here are very similar to the ones from the linear probability model. A big difference here between this model and the linear probability model is that the marginal effects are not constant as they were with the linear probability model. Education has an increasing marginal effect in that its effect on the probability of someone supporting the death penalty increases as education increases. This means that a person completing a graduate degree instead of just a bachelor’s degree (going from 16 to 18 years) is likely to have more of an effect on a person’s views on the death penalty than that person completing high school instead of dropping out early (going from 10 to 12 years). This actually makes more sense and seems more realistic to me than having the marginal effects of education as constant.

* Compare the cross-tabulation of the predicted and actual values from this model to the one you created in question 2. How different are the accuracies of the two prediction models?

A math equation with numbers and words

Description automatically generated with medium confidence

The cross tabulation of predicted and actual values from this model are quite similar to the ones from question 2. This makes sense because the results that were looked at previously between the two models have been quite similar. The linear probability model predicted correctly 59.39% of the time whereas the logit model predicted correctly 59.42% of the time which means the logit model was slightly more accurate with its predictions.

1. (7 points) Create your own model to try and improve our accuracy at predicting who favors/opposes the death penalty. To do this, create some categorical variables for age (at least 3 categories) to test a theory about how this opinion might differ by age. Beyond that, add variables you believe will improve the predictive power of our model.

* Discuss the age categories you are creating, and what you expect to find.

I am adding the age categories under 40, 40 to 60, and over 60. I chose these categories after looking at the frequency distribution for age and judging for 3 roughly equal sized groups. I expect that under 40 will be less in favor of the death penalty than the other two groups, 40 to 60 will be the second group most in favor of the death penalty, and over 60 will be the most in favor of the death penalty. My rationale for this is that people become less merciful as they age because they have experienced the difficulties of life.

* Discuss what other variables you will add (and how, if you are creating dummies, etc.), and what you expect to find.

In addition to adding the previously mentioned age groups and keeping “educ” and “female” as variables in my model which I think will have similar effects as they did in the previous model (am forgoing income as it previously wasn’t statistically significant), I am adding “owns\_gun”, a dummy called “rep” for if a person is a Republican (includes lean Republican, Republican, and strong Republican), and a dummy called “has\_kids” which is if the person has 1 or more kids. I chose to add Republican as a dummy instead of Democrat because I feel like Republican feel more strongly about the death penalty than Democrats. I predict that all three variables will have a positive impact on the probability that someone supports the death penalty.

* Report/copy your results. You may use the LPM or Logit model for this one. Briefly discuss which you chose, and why.

A screenshot of a computer

Description automatically generated

I chose to use the logit model because that way I didn’t have to make the percent predicted correctly table myself, and the logit model is also the more statistically correct model of the two models.

* Summarize the findings of your model. Were your theories correct? What did you find out about the relationship between age and opinion towards the death penalty? Did your model improve in terms of predictive power?

My R^2 is higher than it was in the previous logit model although it is still not as high as I would like it to be. People over 60 had a negative coefficient but were not statistically significant meaning they are not different from people under 40. Having children had a positive coefficient; however, it also was not statistically significant. People aged 40 to 60 had a positive coefficient and was statistically significant which meant they were different from people under 40. Female and education were statistically significant and had negative coefficients. Owns a gun and republican were both quite statistically significant and had quite a high impact on whether or not someone is in favor of the death penalty with them both having positive coefficients. The number of cases correctly predicted was also higher than it was in the previous model. I was mostly correct in my prediction. People over 60 did not have a statistically significant positive coefficient like I was predicting, and although the coefficient for people with children was positive, it was not significant; I was correct with the rest of my prediction though.

* Create/present the “% correctly predicted” table and discuss how this compares to our previous models.

A math equation with numbers and a few words

Description automatically generated with medium confidence

The percent of predictions that my model correctly predicted was higher than the previous model as the new model correctly predicted 66.4% of the cases whereas the previous model correctly predicted 59.4% of the time which is an increase of 7 percentage points.

1. (6 points) For this question, we will work with the variable **quality**, which is a question about how the individual rates the quality of their life (a life satisfaction measure).

* Create an OLS regression model that uses at least four explanatory variables and uses the dependent variable of **quality** as a quantitative variable.

In my model I am including the variables “webhrs”, “attends\_church” which is a variable I created and is true if a person ever attends church, “college” which is a variable that is true if a person has any college education (13 or more years of education), and a dummified categories of the “health” variable.

* Explain your theories about what you expect from the variables you are putting in your model.

I predict “college” will have a positive impact on people’s quality of life because personally my experience in college has helped me to realize what matters most to me in life; I associate this with a higher quality of life. I predict “webhrs” will have a negative coefficient as its very easy to waste time on the internet doing unfulfilling things (I also feel like the news and current events which is kind of covered by this might negatively impact a person as the news tends to be negative), and personally I don’t feel great if I wasted my time. I predict “attends\_church” will have a positive coefficient because I think a lot of people think they have found meaning through religion, and I think meaning leads to a higher quality of life. I am using “poor\_health” as the reference category for the other health variables, and I think the other health variables will have positive coefficient which is greater in magnitude the better the health reported because I would think life is less enjoyable for most if you feel bad health wise all the time.

* Report/copy your results.

A screenshot of a computer

Description automatically generated

* Summarize the findings of the model.

“webhrs” does have a negative coefficient like I predicted; however, it is not statistically significant. All the other variables are quite statistically significant. I was correct about “attends\_church” and “college” having positive coefficients although they are both smaller than I thought they were going to be. Some college education has more of a positive impact on peoples’ quality of life than if the person attends church. I was also correct about the other health variables having a positive coefficient which is greater in magnitude the better the health reported; I was surprised how much this did matter here as “fair\_health”, “good\_health”, and “excellent\_health” all have a greater magnitude than all the rest of the variables. The adjusted R^2 is 0.245933 which means there is still quite a lot of room to improve the model.

* Make up an individual with specific values for each of your variables. You can give this character a name/backstory if you want. Report the values you are plugging in and the predicted quality. Does the estimate seem reasonable?

John is a senior student in college who doesn’t use the internet, attends church twice a year, and is in fair health.

John’s Predicted Quality of Life = (0.286315 \* 1) + (0.143689 \* 1) + (0.496628 \* 1) = 0.926632

Based on this model, John’s quality of life is considered poor. This prediction seems reasonable as John fits the profile of your average engineering student. Despite this, my model might be bit off. Based on it, the only way to have a higher quality of life is by having your health in a better state. There is also no way according to the model to have an excellent quality of life. To make a better model, it would be good to have more variables to capture a very nuanced topic.

1. (7 points) You are going to replicate what you did in question 5, but this time, you will use a binary dependent variable. Create a binary variable from the **quality** variable.

* Discuss your decision as to how to create the binary dependent variable.

I am making the binary dependent variable dependent variable be called “great\_quality” and it is true if the person’s value for quality is very good (4) or excellent (5). I made the decision after looking at the frequency distribution for quality and trying to create two equally sized groups.

* Run a linear probability model with your quality of life binary dependent variable and the same set of explanatory variables as you used in question 5. Report/copy your results.

A screenshot of a computer

Description automatically generated

* Discuss the findings of this model, and how they compare to those from question 5.

Everything that was statistically significant from before is still statistically significant except for the constant. The coefficient for “webhrs” went from barely negative to barely positive in this new model, though it is still not statistically significant. The rest of the findings found from the previous model carry over to this new model with each of the variables carrying about the same level of impact relative to everything else as it did in the model from question 5.

* Predict the y-variable for the same individual as you used in question 5. Does this estimate seem reasonable?

John is a senior student in college who doesn’t use the internet, attends church twice a year, and is in fair health.

Probability John has a great quality of life = (0.156132 \* 1) + (0.0862858 \* 1) + (0.114680 \* 1) = 0.3570978

The model says that there is a 35.71% probability that John considers himself to have a great quality of life. Considering where the calculated value was at with the previous OLS model, I would have guessed this new calculated probability value would have been lower. It is overall saying that it is less likely that John would have a great quality of life than someone who has a 50% probability of a great quality of life; this is what I’d expect so the estimate does seem reasonable.

* Overall, what do you feel is the better approach to looking at the factors impacting quality of life?

Although it might be oversimplifying things, I do think the model with the binary variable might be a better approach because it is simplifying things. I do think it is especially important to simplify things when it comes to asking people an abstract concept like how their quality of life is. Even though having a scale is more nuanced, it requires that people be very good judges of how they truly feel. Lots of choices might overwhelm people or give them more options where they might misjudge how they actually feel; giving them two choices makes it very easy to compare the choices and therefore easier to evaluate how one actually feels. The scale from 1 to 5 also has more room for relativity where people have different understandings of what each number on the scale entails whereas the binary option, although still relative, probably carries a more similar understanding among people than the model with scale from 1 to 5.

* Discuss how you would improve the model predicting quality of life going forward.

To improve the model, I think it would be good to drop “webhrs” from the model as it is not statistically significant. I think adding some new variables to capture more aspects of a nuanced topic would be good. This might entail accounting for variables like how much a person earns in income, what economic class the person is in, how happy the person is, which region the person lives in, whether or not a person is married or divorced, or the age group of the person among other things. Maybe I should have kept the continuous variable of education instead of the dummy variable of college. The adjusted R^2 of the model is 0.184330 so there is quite a lot of room for improvement of the model.