ECON 453
Fall 2023
Problem Set 3 – 38 points
ANSWER KEY

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.

```
Model 1: OLS, using observations 1-3308 (n = 3009)
Missing or incomplete observations dropped: 299
Dependent variable: favor_death

             coefficient    std. error    t-ratio    p-value
  ------------------------------------------------------------
  const       1.04740       0.0591256     17.71     7.23e-067 ***
  age         0.00116178    0.000538222    2.159    0.0310    **
  educ       -0.0356385     0.00358436    -9.943    6.11e-023 ***
  female     -0.0799075     0.0180775     -4.420    1.02e-05  ***
  income      0.000240759   0.000182364    1.320    0.1869

Mean dependent var   0.549352   S.D. dependent var   0.497641
Sum squared resid    714.2865   S.E. of regression   0.487625
R-squared            0.041125   Adjusted R-squared   0.039848
F(4, 3004)           32.20933   P-value(F)           2.54e-26
Log-likelihood      -2105.996   Akaike criterion     4221.993
Schwarz criterion    4252.039   Hannan-Quinn         4232.799
```

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

The female coefficient tells us that females, all else equal, are about 8 percentage points less likely than males to favor the death penalty as punishment. Every year of education drops the likelihood a person favors the death penalty by about 3.6 percentage points. The age coefficient says there is a slight increase in probability of favoring death penalty as people age. Age, education, and gender are all statistically significant. Income has a positive coefficient but is not estimated to have a significant relationship with death penalty views. In terms of matching expectations, how would I know what your individual expectations are? Personally, the education and female coefficients seem reasonable to me. I do not think age would have a linear relationship (see below), so not sure what to make of that one. I do not see why income would factor into a person's views, so the lack of significance doesn't surprise me.

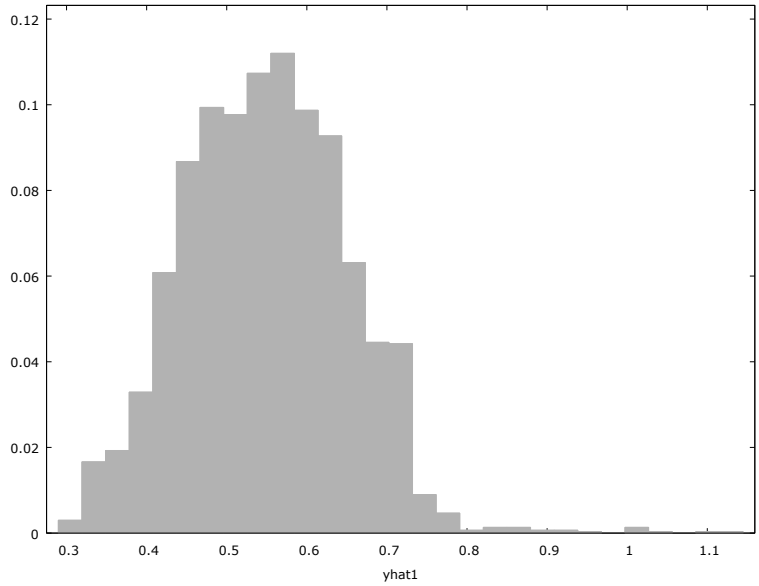
   - Predict the probability that a 75-year-old male with an income of $60,000 per year is in favor of the death penalty. Compute this for 5 different levels of education: 10, 12, 14, 16, and 18 years. Report the probabilities and comment briefly if this seems like a reasonable set of marginal effects of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

I used Excel to do these, here are my calculations:

Age	75	75	75	75	75
Education	10	12	14	16	18
Female	0	0	0	0	0
Income	60	60	60	60	60
Pred Probability	79.3%	72.1%	65.0%	57.9%	50.7%
Change		-0.07128	-0.07128	-0.07128	-0.07128

To me, the probabilities are reasonable estimates. We know the value is around 55% favor for the overall sample. This estimate is for a male, which we know will have higher probabilitie. In terms of whether the marginal effects are reasonable, that is a little bit trickier. Since this is a linear model, we are getting estimates that say every 2 years of education decrease the likelihood of favoring the death penalty by the same amount. Going from 10 to 12 years has the same effect as going from 12 to 14, and so on. This is probably not what I would expect, since, for example, going from 14 to 16 usually means completing a bachelor's degree. I would think that might impact someone's views differently than, for example, completing their high school degree.

2. (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.



We learn that most of the predicted values fall in a reasonable range of about 30 to 70 % chance of favoring the death penalty. There are only a handful of observations that give us unreasonable/nonsensical predictions. No one is predicted to have a negative probability, but there are a few with a predicted probability of over 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.

```
Cross-tabulation of favor_death (rows) against predict_favors (columns)

       [   0][   1]  TOT.

[   0]    568   788  1356
[   1]    434  1219  1653

TOTAL    1002  2007  3009

299 missing values

Pearson chi-square test = 81.9602 (1 df, p-value = 1.38858e-019)
```

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

From the table, we can see that our model predicted correctly for 1219 people that favored the death penalty and for 568 that did not. This means we correctly predicted 59.4% of the sample ((1219 + 568) / 3009). The larger "mistake" was predicting people favored the death penalty when they, in fact, do not. This happened for 788 people. There were 434 people that our model predicted did not favor the death penalty when they, in fact, did.

3. (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.

```
Model 2: Logit, using observations 1-3308 (n = 3009)
Missing or incomplete observations dropped: 299
Dependent variable: favor_death
Standard errors based on Hessian

             coefficient   std. error      z         slope
  ------------------------------------------------------------
  const       2.35042      0.260401       9.026
  age         0.00482400   0.00226506     2.130     0.00119300
  educ       -0.153011     0.0158769     -9.637    -0.0378403
  female     -0.336799     0.0763394     -4.412    -0.0829633
  income      0.00105178   0.000766765    1.372     0.000260111

Mean dependent var   0.549352   S.D. dependent var   0.497641
McFadden R-squared   0.030790   Adjusted R-squared   0.028376
Log-likelihood      -2007.232   Akaike criterion     4024.463
Schwarz criterion    4054.510   Hannan-Quinn         4035.269

Number of cases 'correctly predicted' = 1788 (59.4%)
f(beta'x) at mean of independent vars = 0.247
Likelihood ratio test: Chi-square(4) = 127.534 [0.0000]

           Predicted
               0      1
  Actual 0   571    785
         1   436   1217

Excluding the constant, p-value was highest for variable 7 (income)
```

   - 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?

For interpretation purposes, we want to look at the "slope" portion of the Logit regression results. These are remarkably similar to the coefficients in our linear probability model. The female coefficient, for example, was 8.0 percentage points in the LPM and is 8.3 percentage points in the Logit. There are really no major differences, the magnitude, sign, and significance of each of the variables are all similar across the two models. This is often what we see, which is why people often tend to rely on the simpler-to-interpret LPM results.

   - 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.

For this one, you need to get the exponential function involved. Here are my estimates from Excel:

Age	75	75	75	75	75
Education	10	12	14	16	18
Female	0	0	0	0	0
Income	60	60	60	60	60
EXP	3.473688	2.557921	1.883577	1.387011	1.021354
Pred Probability	77.65%	71.89%	65.32%	58.11%	50.53%
		-5.8%	-6.6%	-7.2%	-7.6%

The levels of the predicted probabilities are very similar to those from the linear probability model. What you should notice is that the marginal effects of education are not linear for this one. In particular, there is an increasing effect of education on the likelihood of favoring the death penalty. Going from 10 to 12 years of education does not have as large an impact on people's views as going from 12 to 14, and so on. I actually think this makes more sense; higher levels of education probably have more of an impact on a person's political/societal views.

- 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?

The nice thing about the Logit model is that it gives you the "% correctly predicted" table automatically. If is part of the results we saw earlier in the question.

```
Number of cases 'correctly predicted' = 1788 (59.4%)
f(beta'x) at mean of independent vars = 0.247
Likelihood ratio test: Chi-square(4) = 127.534 [0.0000]

           Predicted
               0      1
  Actual 0   571    785
         1   436   1217
```

The numbers here are very similar to what we calculated for the linear probability model in question 2. This shouldn't be surprising given the similarities in the results of the two models that we have already discussed. Our LPM correctly predicted 59.4% of cases, and the Logit correctly predicts for 59.4% of cases.

4. (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 chose to make 4 categories: under 30, 30 to 45, 46 to 65, and over 65. My expectations are that young people will be less likely to favor the death penalty (they are less likely to be jaded by life experiences), middle-aged people will be the most in favor, and older people will be somewhere in between (people's views might soften as they age). I am not sure what to expect between the two "middle-aged" categories I have, I just wanted to try it. Let me have my fun, guys.

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

I kept the education and female variables from earlier and dropped the income variable. I decided to add the variables about gun ownership and political affiliation. I included "own_gun" and made a dummy for those that identified as Republican (I did not include leaning Republican in this group). I am expecting each of these to positively impact the likelihood of favoring 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.

I chose the Logit model because I like to do things that are statistically correct but mostly because I didn't want to create the % correctly predicted table (later in the question) by hand.

```
Model 4: Logit, using observations 1-3286 (n = 3235)
Missing or incomplete observations dropped: 51
Dependent variable: favor_death
Standard errors based on Hessian

              coefficient   std. error     z        slope
  ------------------------------------------------------------
  const          1.42551     0.239853      5.943
  educ          -0.127798    0.0141871    -9.008   -0.0314373
  female        -0.223360    0.0768977    -2.905   -0.0548057
  age30_45       0.394224    0.131122      3.007    0.0954838
  age46_65       0.517839    0.128392      4.033    0.125344
  age_over65     0.149671    0.134937      1.109    0.0366155
  owns_gun       0.645816    0.0822781     7.849    0.155499
  repub          1.31998     0.105937     12.46     0.292219

Mean dependent var   0.549304   S.D. dependent var   0.497640
McFadden R-squared   0.096785   Adjusted R-squared   0.093193
Log-likelihood      -2011.077   Akaike criterion     4038.154
Schwarz criterion    4086.808   Hannan-Quinn         4055.588
```

- 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?

According to my results, my theories were mostly correct. People under 30 to 45 and 46 to 65 are more likely than people under 30 to favor the death penalty. The coefficient on my old people variable (over 65) is positive, but is not significant, so the under 30 and over 65 categories do not differ. Education and female have the same effects as before (negative, and significant). The gun ownership and Republican variables are both positive and highly significant, as I expected. My adjusted R^2 is higher than what we have seen in other models, though still relatively low in the grand scheme of things.

- Create/present the "% correctly predicted" table and discuss how this compares to our previous models.

Since I was smart and estimated a logit model, this table was already created as part of my output. The % correctly predicted did increase in my model. Earlier, we were correctly predicting for about 59% of cases. Now, my value is up to 64.8%. Maybe this is just the start for a new, more confident Dan that accomplishes great things.

```
Number of cases 'correctly predicted' = 2096 (64.8%)
f(beta'x) at mean of independent vars = 0.246
Likelihood ratio test: Chi-square(7) = 431.001 [0.0000]

           Predicted
               0      1
  Actual 0   900    558
         1   581   1196

Excluding the constant, p-value was highest for variable 24 (age_over65)
```

5. (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.
   - Explain your theories about what you expect from the variables you are putting in your model.

I included female to begin with. I don't really have any expectations for this one, just curious. Then I included the age categories that I used in question 4. I think younger (under 30) and older (over 65) will be more satisfied with life than middle-aged people. I included the religion variable, and am guessing that the more religious someone is, the more satisfied they are with life. Next, I made a dummy for whether someone has at least a bachelor's degree. I would expect this has a strong positive impact on quality of life (for income, job satisfaction, and other reasons). Finally, I included a couple of categories for marriage. In particular, I made dummies for those that are currently married (marst==1) and those

that have never been married (marst==5). I would guess these two groups are happier than those that are in my reference category (divorced, widowed, and separated).

- Report/copy your results.

```
Model 5: OLS, using observations 1-3308 (n = 3287)
Missing or incomplete observations dropped: 21
Dependent variable: quality

                coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const          2.96621      0.0697401    42.53     0.0000    ***
  female        -0.0109814    0.0299416    -0.3668   0.7138
  college        0.513627     0.0297820    17.25     7.25e-064 ***
  religion       0.0583719    0.0141537     4.124    3.81e-05  ***
  married        0.392751     0.0370989    10.59     8.92e-026 ***
  never_married  0.0366294    0.0472326     0.7755   0.4381
  age30_45      -0.115070     0.0546272    -2.106    0.0352    **
  age46_65      -0.0286478    0.0563437    -0.5084   0.6112
  age_over65     0.112479     0.0598297     1.880    0.0602    *

Mean dependent var   3.520840   S.D. dependent var   0.912869
Sum squared resid    2314.780   S.E. of regression   0.840331
R-squared            0.154672   Adjusted R-squared   0.152609
F(8, 3278)           74.97318   P-value(F)           6.8e-114
Log-likelihood      -4087.740   Akaike criterion     8193.481
Schwarz criterion    8248.360   Hannan-Quinn         8213.129
```

- Summarize the findings of the model.

Overall, the explanatory power of my model is not tremendous, with an adjusted R^2 of only about 0.153. The female variable was insignificant, which did not surprise me. I found that 30 to 45-year-olds report a lower quality of life than those under 30. The 46 to 65 age group, however, has no difference in quality from the young people. Good news, folks, just a few years and my quality of life will skyrocket. Older people have higher quality, though there is only weak significance for this estimate. The college variable is positive, significant, and large in magnitude. Married people report a higher quality of life than divorced/separated/widowed, but there is no difference between the latter group and those that have never been married. The more religious people are (measured by attendance), the higher their self-reported quality of life (and their chances of having a successful afterlife experience).

- 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?

I will be predicting the values for Chuck Norris. Chuck is an 83-year-old married male without a college degree. According to Wikipedia, an "…outspoken Christian, Norris is the author of several Christian-themed books." Depending on how much you want to know, I could also tell you the specific name of the church he attends. In any case, I will estimate his religion value as 4 (at least once a week). My estimate for Chuck Norris's quality of life:

Predicted quality = 2.966 + (0.0584*4) + (0.393*1) + (0.112*1) = 3.7046.

According to my model, Chuck Norris somewhere between good and very good in terms of his quality of life, leaning towards the very good.

6. (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 decided to create a variable called "high quality" that is 1 if the person is a 4 or 5 on the quality-of-life scale (very good or excellent). I decided this because I looked at the frequency distribution of the quality variable, and more than half of the sample (about 53%) fall in these two categories.

- 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.

```
Model 6: OLS, using observations 1-3308 (n = 3287)
Missing or incomplete observations dropped: 21
Dependent variable: high_quality

                coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const          0.272017     0.0388618     7.000    3.10e-012 ***
  female        -0.0106432    0.0166846    -0.6379   0.5236
  college        0.257054     0.0165957    15.49     2.80e-052 ***
  religion       0.0376015    0.00788699    4.768    1.95e-06  ***
  married        0.170997     0.0206729     8.272    1.90e-016 ***
  never_married  0.0199725    0.0263198     0.7588   0.4480
  age30_45      -0.0853107    0.0304403    -2.803    0.0051    ***
  age46_65      -0.0341735    0.0313968    -1.088    0.2765
  age_over65     0.0337343    0.0333393     1.012    0.3117

Mean dependent var   0.532400   S.D. dependent var   0.499025
Sum squared resid    718.7710   S.E. of regression   0.468264
R-squared            0.121628   Adjusted R-squared   0.119485
F(8, 3278)           56.73817   P-value(F)           6.56e-87
Log-likelihood      -2165.622   Akaike criterion     4349.245
Schwarz criterion    4404.124   Hannan-Quinn         4368.893
```

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

The general nature of the findings in terms of sign/significance are very similar to those in question 5. The old people category was weakly significant in the previous model and is not at all significant in this one. Other findings, such as the large positive effects of marriage and college education, are consistent.

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

Let's estimate for Chuck Norris again:

Predicted probability of high quality = 0.272 + (0.0376*4) + (0.171*1) + (0.0337*1) = 0.6271 = 62.7%.

My model estimates about a 62.7% chance that Chuck Norris considers himself to have a high quality of life. That makes me kind of sad, given all that he has accomplished. If Chuck Norris is not satisfied with his quality of life, what chance is there for those of us struggling to get our orange belts.

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

This is a tricky variable to deal with, and there is not a "correct" answer to this question. The ladder from 1 to 5 gives us some detail about where a person rates their quality, but it is also a little bit odd to think about as a scale. Is this difference between fair and poor the same as the difference between very good and excellent, for example? This is what we are implying if we use the quality scale as a quantitative dependent variable. The binary option (used in question 6) does not suffer from the scaling problem, but it might be oversimplifying things. What two categories should be used in this case? I chose to estimate the probability a person reports a 4 or 5, but you might have grouped 3, 4, and 5 together, or perhaps to estimate the probability that the person is a 5.

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

One thing that I would like to noodle around with is adjusting my age categories to see if that makes any difference. The way they are currently defined showed some differences, but it seems like there is room for improvement there. I should probably think about incorporating income or "class" into the model as well, this seems like an important part of how people evaluate their lives. Maybe it would be better to use the continuous years of education variable instead of the dummy for having at least a bachelor's degree? That is another thing I would play around with. Overall, I tried my best, but there is always room to reflect and think about what we might do differently next time.