**Group 23**

Ishaan Dey

Jedidiah Park

Angela Li

Austin Rule

Lab #10

4/17/19

ECON 3720

University of Virginia

Ron Michener

*Lab #10*

This assignment is an exercise in estimating probabilities using both the linear probability model, logit, and probit. You will gain experience in how to handle a dummy dependent variable.

The data set contains observations on several variables for all 120 counties in Kentucky. You will use the data to estimate the probabilities that counties will choose to be wet or dry; that is, whether they will allow the legal sales of alcoholic beverages. The data were gathered in the 1990s and many of the variables were measured in 1990. It is helpful to know that among major denominations, Baptists tend to be the least tolerant of alcohol, while Catholics tend to be among the most permissive – the difference is likely cultural and historical rather than theological, but that is another story.

Weti = A dummy variable that is equal to one if county i is wet, zero if otherwise.

Baptisti = The percentage of the population of county i that is affiliated with the Baptist church, including Southern and United Baptist churches.

Catholici = The percentage of the population in county i that is a member of the Roman Catholic church.

Percent\_nonwhitei = The percentage of the population in county i that is African-American plus the percentage of the total population of county i that is Hispanic.

RUC\_codei = The USDA rural-urban continuum code for the county. Counties are coded on degree of urbanization, where zero is the most urban and nine is the least urban. A full description is available here. <http://www.ers.usda.gov/briefing/Rurality/RuralUrbCon/>

Step I.

**(5 points) You are to fit a linear probability model explaining wet with all the other variables as explanatory variables, treating the variable RUC\_code as a quantitative variable. What signs do you expect to get? Fit the model with OLS and include the results in your report. Sp9a**

BBaptist: Negative, as Baptists are intolerant of alcohol

BCatholic: Positive, as Catholics are permissive of alcohol

BPercentNonWhite: Ambiguous, as it is uncertain how ethnic minorities differ in alcohol preference

BRUC\_Code: Ambiguous, as it is uncertain how urban/rurality affects alcohol perception



Step II.

**(4 points) Use the linear probability model to generate the predicted probabilities of a county being Wet. How many of these probabilities are greater than one? Less than zero?**

**Sp9a**



8 Counties have a predicted probability less than 0, and 7 with greater than 1.

Step III.

**(6 points) Using a 5% significance level, test for heteroskedasticity in the linear probability model estimated in Step I in two ways: Use a Breusch-Pagan test, and use the White test. Include the results in your report. Do you conclude you do or don’t have heteroskedasticity? Are the results surprising? Explain why or why not. Sp9a**





With a p value less than alpha (0.0273 and 0.0060 for White and Breush-Pagan, respectively), we can conclude that the model exhibits heteroskedasticty. This is not surprising, as there are obvious differences in variance between the different conditions of the dependent variable.

Step IV.

**(6 points) HC corrected standard errors are designed to address the problem of pure hetroskedasticity, and the linear probability model suffers from other specification errors. However, if one is going to estimate the linear probability model, it is arguably more sensible to use HC standard errors. Estimate the linear probability model again using HC3 standard errors (those are the ones with bias correction of). Include the result in your report. What are the p-values associated with the one-sided hypotheses you set forth in the first step? Which variables are statistically significant at the 5% level when using HC3 standard errors? Sp9a**



Baptist p-value = 0.000

Catholic p-value = 0.001

Percent\_nonwhite p-value = .0015

RUC\_code p-value = o.000  
\*All variables are statistically significant at alpha = 0.05

Step V.

**(6 points) Using the rule that you predict “Wet=1” if the predicted probability is more than 0.5 and that you predict “Wet=0” if the predicted probability if 0.5 or less, generate predicted probabilities of a county being Wet using the specification fit in Step I. Use them results to compute Studenmund’s measure (see p. 394) Sp9a Sp9a**



R2p = 0.5(69/(69+6)+27/(18+27))\*100 = .7600

Step VI.

(**6 points) Now fit the same model using logit; that is . Once again treat RUC\_code as a quantitative variable. Include a copy of the output in your report. Do all the variables have the predicted signs? What are the p-values associated with each variable and which are statistically significant at the 5% level? Sp9a**



All of the coefficients have the expected signs and are statistically significant at the 5% level.

Baptist p-value = 0.0005 (significant)

Catholic p-value = 0.0225 (significant)

Percent nonwhite p-value = 0.0035 (significant)

RUC Code p-value = 0.0005 (significant)

Step VII.

**(6 points) How many of the dry counties are correctly predicted by the logit model? How many of the Wet counties are correctly predicted? Include the relevant STATA output in your results. What is Studenmund’s** **measure? Sp9a**



Studenmund’s R2p measures the average of the true positive and true negative rates. The value for this model is .7889

Step VIII..

**(5 points) So far we have treated the RUC\_code as if it is a quantitative variable, but it could be argued that it should really be modeled with a set of dummy variables. The difficulty with doing so is that it will require many dummy variables, eating up many degrees of freedom, and our sample size is not very large. Entering the RUC\_code as a single quantitative variable is a simplification that makes sense when you have a small sample. In a very large sample, it would almost certainly make more sense to define a set of dummy variables instead. Whether 120 observations is a “small” or “big” sample, however, is debatable, and one might want to see which works best from a goodness of fit perspective. Use variable definition commands to create a set of dummy variables for the RUC\_code and then use the dummies in place of the RUC\_code in a logit model to explain Wet/dry (all other variables the same). Include the result in your report. Sp9a**



Step IX.

**(4 points) Test to see whether all the dummy variables might simultaneously have zero coefficients in the logit regression. The STATA command works just as in OLS. Do you accept or reject the null using an alpha of 5%? Sp9a**



p(0.1024) > alpha(0.05) thus fail to reject the null hypothesis that the variables might simultaneously have zero coefficients.

Step X.

**(6 points) Another way to approach the problem is to ask whether the Step VI specification or the Step VIII specification fits best. Which version has the higher pseudo-R-squared? Is this helpful? Logit and probit specifications can be compared based on their AIC and BIC scores, just as we did for OLS regressions, although the numerical calculations are different and equations (6.26) and (6.27) can’t be used. Review the principles underlying the AIC and BIC criterion on pp. 187-88. Compare the two specifications this way (the STATA command works just as it did in OLS). Which version is preferred by this measure? Include the STATA output in your report. Sp9a**

The i.RUC model has a higher R2 (.3690) than the c.RUC model (.3414).

For logit w/ i.RUC:



For logit w/ c.RUC:



We thus prefer the c.RUC model based upon the information criterion as both measures are lower (114.6 < 124.2 & 128.5 < 157.6).

Step XI.

**(4 points) Using probit, that is, , fit the model including all the other explanatory variables. Treat RUC\_code as a quantitative variable. Include a copy in your report. Sp9a**



Step XII.

**(8 points) How many dry counties does the probit model fit in Step XI correctly predict? Wet counties? What is the goodness of fit measure, ? Include the relevant STATA output in your report. Compute the goodness of fit of the models fit in Steps I, VI, and XI. Which seems to do the best job? Sp9a**

Classification for probit model: ****

The model correctly predicts 71 of 75 dry counties.

R2p of Probit = .5(71/75 + 27/45) = .7733

R2p of Logit = .5(70/75 + 29/45) = .7889

R2p of LPM = .5(69/75 + 27/45) =.7600

The logit model appears to be the best based upon R2p values. The probit model is also better for both AIC and BIC, although only slightly relative to logit. Both of these models, however, are still favored far more than the linear probability model.

IC for Probit:



IC for Logit:



IC for LPM:



.

Step XIII.

**(14 points) Now compare the marginal impact of a change in each of the independent variables in a table as predicted by each of the models (fit in Step I, Step VI, and Step XI). For the logit and probit models, evaluate the marginal impact in three ways – according to Studenmund’s rule of thumb for translating probit and logit coefficients into their linear probability analogues, by considering the marginal impact for an average county (that is, a county with all explanatory variables at their means), and finally by considering the average marginal impact. Present the results in a table like that below, and include the STATA output needed to evaluate the marginal impacts in the logit and probit models. Sp9a**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| S19a | Baptist | Catholic | Percent\_nonwhite | RUC\_code |
| Linear model | -.0073163 | .0123618 | .0247808 | -.0556068 |
| Logit, approximation | -0.0150622 | 0.0287369 | 0.046626525 | -0.07873545 |
| Logit, at county means | -.0137983 | .0263255 | .042714 | -.0721285 |
| Logit, average marginal | -.0085423 | .0162978 | .0264437 | -.0446539 |
| Probit, approximation | -0.01384732 | 0.02693336 | 0.04143536 | -0.07336868 |
| Probit, at county means | -.0129183 | .0251264 | .0386555 | -.0684465 |
| Probit, average marginal | -.0084388 | .0164137 | .0252515 | -.0447123 |

Linear Probability Model:



Logit Models (County Means, Average marginal)





Probit Models (County mean, Average marginal)





Step XIV.

**(5 points) Use the logit model fit in step VI to estimate the probability that a county with the following characteristics will be Wet:**

Baptisti = 15

Catholici = 4

Percent\_nonwhitei = 5

RUC\_codei =4

**Include the output in your report. Sp9a**



OR



Probability of Wet = .7492686

Step XV.

**(5 points) Repeat Step XIV if Catholici = 50 (all other variables remaining the same). Sp9a**



OR



Probability of Wet = .9983116

Step XVI.

**(10 points) Compute the predicted marginal impacts of Baptist, Catholic, Percent\_nonwhite, and RUC\_code estimated at the values given in Step XV with those estimated at the values given in Step XIV. Include the estimated marginal impacts in your report. Confirm that the marginal impacts are much smaller in Step XV than in Step XIV. Explain why.**

The marginal impacts at Catholic = 4 are much larger than when Catholic = 50 (0.0216 > 0.000194). This is because as the observation moves closer to the ends of the variable range, the probability curve plateaus such that it does not make as much of an increase in probability for a per unit change in Catholic percentage in that county.  **acc9a**

Ge

