Joel Cabrera

Econometrics (01:220:320)

Professor Agan

2 May 2019

**R Data Analysis - HW #2**

Part 1: Presidential Voting

1. & 1a. (See R code for new categorical variable).

2. (see R code for probit regression of voting Republican on the high-income variable and the marginal effect at the mean of the independent variables).

2a. Probit is used here because not only does the regression model deal with categorical data – that is, it contains a binary independent and a binary dependent variable, but also the relationship between the predicted probability of voting Republican and the high-income predictor is not linear. Marginal effects are reported, and not the direct coefficients, because the former allows for an easier, better, and more intuitive interpretation than the latter. Interpreting the former allows you to understand an independent variable’s percentage point change, which affects the probability of the dependent variable occurring. Interpreting the latter, however, only yields an understanding of a change in the z-score.

2b. Having a high income is associated with a 12.7 percentage point increase in the probability of voting Republican in the 2000 U.S. presidential election.

3. For age: a 1-year increase in age is associated with a 0.3 percentage point increase in the probability of voting Republication in the 2000 U.S. presidential election, holding all other binary independent variables at 0.

For female: being female is associated with a 9.5 percentage point decrease in the probability of voting Republication in the 2000 U.S. presidential election, holding age at its mean and all other binary independent variables at 0.

For race, blacks and whites, respectively: being black, as compared to other races, is associated with a 27.3 percentage point decrease in the probability of voting Republication in the 2000 U.S. presidential election measured, holding age at its mean and all other binary independent variables at 0. Being white, as compared to other races, is associated with a 10.9 percentage point increase in the probability of voting Republication in the 2000 U.S. presidential election, holding age at its mean and all other binary independent variables at 0.

The marginal effects of the age, female, white, and black variables on voting Republican in the 2000 U.S. presidential election, assuming a 5% level is used, are statistically significant, except for the female and black ones. These tests of significance were done by dividing each variable’s marginal effect by 2. If the result was greater than its corresponding standard error, that variable’s marginal effect was statistically significant at the 5% level.

4. Another (binary) independent variable that could be added to the probit regression model would be college. Dealing with omitted variable bias: college attendance could have been a determinant of voting Republican in the 2000 presidential election in that, for example, being college educated would be associated with a lower probability of voting for Republican. Being white could be correlated with going to college in that, for example, the former is associated with a higher probability of doing the latter. Interpreting the marginal effect of college attendance on voting Republican: going to college is associated with a 0.9 percentage point increase in the probability of voting Republication in the 2000 U.S. presidential election, holding age at its mean and all other binary independent variables at 0.

**Table 1: Determinants of Voting Republican in the 2000 U.S. Presidential Election**

|  |  |  |  |
| --- | --- | --- | --- |
|  | (1) | (2) | (3) |
| High-Income | 0.127\*\*\*  (0.031) | 0.078\*  (0.031) | 0.075\*  (0.033) |
| Age |  | 0.003\*  (0.001) | 0.003\*  (0.001) |
| Female |  | -0.095\*\*  (0.030) | -0.095\*\*  (0.030) |
| Black |  | -0.273\*\*\*  (0.037) | -0.273\*\*\*  (0.037) |
| White |  | 0.109\*\*  (0.040) | 0.108\*\*  (0.041) |
| College |  |  | 0.009  (0.033) |
| Intercept | -0.589\*\*\*  (0.053) | -0.893\*\*\*  (0.193) | -0.898\*\*\*  (0.194) |
|  |  |  |  |
| Number of Observations | 1015 | 1015 | 1015 |

**Note:** Standard errors in parentheses. \*\*\* indicates p < 0.01. \*\* indicates p < 0.05. \* indicates p < 0.10. Numbers are rounded to 3 digits. Values above standard errors, except for intercepts, are marginal effects, not coefficients. Standard errors of constants are from probit, not marginal effects. High-income level is a categorical level (1 = family income is above the 68th percentile, 0 = otherwise). Other race is the reference group (omitted category). Data is from the votingdata.csv dataset.

Part 2: Determinants of the State Suicide Rate

1. (See R code for regression of adult suicide rate on the independent variables).

1a. A 1 percentage point increase in the state unemployment rate is associated with a 0.264 percentage point decrease in the adult suicide rate between 1990-2000.

1b. Yes, I am surprised by the sign of the coefficient. I expected to be positive instead, as, I would like to think, one who is unemployed is more likely to commit suicide due to having no job or source of disposable income.

2. (See R code for the repeated regression but includes both state and year fixed effects).

2a. The coefficients that changed sign and significance from the first regression are those on: state unemployment rate, state population in rural areas, state population with a college degree, and state population with no insurance. Per capita state did not change in sign but did so in terms of significance level.

2b. Possible explanations for the changes in the above mentioned four coefficients, with respect to state and year fixed effects, are: real state GDP per capita, state incarceration rate, and state domestic violence rate (as well as other independent variables from the dataset that were not used in this 2nd regression). Focusing on the first potential omitted variable, it varies across states and time; it is not constant. It could also be a determinant of adult suicide rates in that, for example, a lower real state GDP per capita is associated with a higher suicide rate. Furthermore, state population with a college degree, one of the independent variables, could be correlated with the real state GDP per capita in that, for example, a higher state population with a college degree – measured in percentage - is associated with a higher real state GDP per capita. After, equipping workers with more skills can increase productivity in the workforce, and thus, allow their businesses to earn more economic profits. Such an omitted variable, then, could have changed the sign and significance of the coefficient on the state population with a college degree.

3a. A 1 percentage point increase in the state Southern Baptist population is associated with a 0.516 percentage point decrease in the adult suicide rate between 1990-2000, controlling for all other independent variables and state and year fixed effects.

3b. Using the “linearHypothesis” command on the religious variables, I received an F-statistic value of 2.9097, and its associated p-value is 0.02127. Because the F-statistic is greater than 1.96, the critical value of the 5% significance level (alternatively, because the p-value is lower than the 5% significance level), we reject the H0, and therefore find the religious variables to be jointly significant.

4. Another independent variable that could be added to the fixed effects regression model, would be the state black population. Dealing omitted variable bias: the state black population could be a determinant of the adult suicide rate between 1990-2000 in that, for example, as the former increases, the latter also increases. The state unemployment rate, for example, could be correlated with the state black population in that, as the former increases, the latter also increases. Interpreting the coefficient on the state black population: a 1 percentage point increase in the state black population is associated with a 0.011 percentage point increase in the adult suicide rate between 1990-2000, controlling for all other independent variables and state and year fixed effects. From the fourth fixed effects regression model (of which its results are reported in Column 4 of Table 2), I believe that this effect that I estimated is not causal. While the fixed effects regression model controls for state and time fixed effects, there are still things that vary across time and states that can cause omitted variable bias. However, yes, we would have more confidence in the causal interpretation than if we ran a cross-sectional regression that used data in a single year only (despite it not being completely correct). As just mentioned earlier, the regression model controls for state and year fixed effects, but we have not eliminated omitted variable bias in doing so. A cross-sectional regression that used data in a single year, however, not only does not explicitly control for state and year fixed effects, but also could lead to more biased results than would rise from a panel-data fixed effects regression (as omitted variable bias would not be really mitigated). Determining causality, while there are many ways to do so, generally involves all the possible existing independent variables that truly and completely determine the dependent variable, thereby eliminating omitted variable bias. (ideally, all such variables are done in an ideal randomized experiment).

5. Yes, there is reason to believe that these heteroskedasticity-robust standard errors may be incorrect. The error terms in the associated fixed effects regression model may be autocorrelated – that is, they may be correlated over time within a state. Because of this autocorrelation, the usual standard errors are no longer valid; they are derived under the assumption that there is no serial correlation. As such, clustered standard errors are needed, which allow error terms to have correlation within a cluster (e.g. entity) but uncorrelated across clusters.

Running the clustered standard errors command in R with respect to the use fixed errors regression model, it is shown, as reported in Column 5 of Table 2, that all the said model’s standard errors have all changed (but not their corresponding coefficients).

**Table 2: Estimated Effects of Demographic Variables on Adult Suicide Rates from 1990-2000**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | (1) | (2) | (3) | (4) | (5) |
| State Unemployment Rate | -0.264\*  (0.111) | 0.126  (0.111) | 0.174  (0.112) | 0.174  (0.112) | 0.174  (0.132) |
| Per Capita State Income | -0.078\*\*\*  (0.012) | -0.067\*\*  (0.022) | -0.053\*  (0.022) | -0.053\*  (0.022) | -0.053\*  (0.02) |
| State Population in Rural Areas | -0.032\*  (0.014) | 0.253\*\*\*  (0.064) | 0.216\*\*\*  (0.065) | 0.216\*\*  (0.065) | 0.216\*\*  (0.083) |
| State Population with 4-Year Degree | 0.041  (0.051) | -0.120\*  (0.052) | -0.090\*  (0.053) | -0.090\*  (0.053) | -0.090  (0.072) |
| State Population with No Insurance | 0.294\*\*\*  (0.044) | -0.008  (0.044) | -0.036  (0.045) | -0.036  (0.045) | -0.036  (0.059) |
| State Mormon Population |  |  | 0.367  (0.260) | 0.367  (0.260) | 0.367  (0.269) |
| State Southern Baptist Population |  |  | -0.516\*  (0.260) | -0.518\*  (0.267) | -0.516  (0.459) |
| State Roman Catholic Population |  |  | -0.030  (0.081) | -0.030  (0.081) | -0.030  (0.109) |
| State Protestant Population |  |  | 0.238\*\*\*  (0.071) | 0.239\*\*  (0.074) | 0.238\*  (0.142) |
| State Black Population |  |  |  | 0.011  (0.293) |  |
| Intercept | 25.202\*\*\*  (2.206) |  |  |  |  |
|  |  |  |  |  |  |
| Number of Observations | 561 | 561 | 561 | 561 | 561 |
| Adj R2 | 0.244 | -0.030 | -0.014 | -0.016 |  |
| State Fixed Effects | No | Yes | Yes | Yes | Yes |
| Year Fixed Effects | No | Yes | Yes | Yes | Yes |

**Note:** Heteroskedasticity-robust standard errors in parentheses. \*\*\* indicates p < 0.01. \*\* indicates p < 0.05. \* indicates p < 0.10. Numbers are rounded to 3 digits. Independent variables, except for per capita state income, are measured in terms of percentage. Adult suicide rate is measured in 100,000s. R2 is based on the adjusted R-squared computed result. Data is from the Suicide.csv dataset.