Kellen Bullock

Dr. Comer

2-29-2020

Exercise 4

1. My assigned variables were Pct\_White, Black\_log10, Hispanic\_log10, Median home value, poverty, SNAP, FIRE\_I, blue collar occupations, unemployed, and two or more races.
   1. This enter model that was run is not very effective at explaining how my assigned variables influence percent of population registered republican. The basis for this conclusion is a R square value of on 0.610 and a standard error of 10.210%. What also says this model is not very good is none of the variables were significant.
   2. This model does not seem to have very much mulitcolinearity, because the variance inflation factor (VIF) for all variables was lower than 5. The two variables I suspect have a stronger multicolinearty issue that what was presented in this model are percent poverty and percent unemployed. Both of their VIFs are 3.004 and 3.905 respectively. Those values are not at the rule of thumb 5 VIF but their both closely related and coincidentally have similar VIFs. There is possible something between percent tow or more races and either Hispanic log10 or Black log10. Their VIFs are 2.241, 1.720 and 1.912 respectively.
   3. The following variables will likely be in the stepwise regression model are: percent white, poverty, and median home value. Percent poverty has a high significance in this model and generally explains unemployment, and SNAP. Percent white does not have a high significance but is a large majority of the population and may have more significance when another variable is dropped. Median home value has a fairly high significance in this model and is also can show how wealth influences voter registration. I do not believe the other variables will make it into the model because they are either explaining each other such as two or more races and Hispanic log10, or they are a proxy to explain income.
2. 1. Step #, variable entered/removed, total R^2, contribution of R^2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step # 1 | Variable | Entered/removed | Total R^2 | Contribution R^2 |
| Percent living in poverty | Entered | 0.479 | -0.692 |
| Black Transformed log10 | Removed |  | -0.103 |
| Hispanic Transformed log10 | Removed |  | 0.147 |
| Median home value | Removed |  | 0.111 |
| Receiving SNAP assistance | Removed |  | -0.355 |
| Percent FIRE\_I | Removed |  | -0.24 |
| Blue Collar Occupations | Removed |  | -0.021 |
|  | Workforce Unemployed | Removed |  | -.221 |
|  | Two or More Races | Removed |  | -0.255 |
|  | White | Removed |  | 0.213 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step # 2 | Variable | Entered/removed | Total R^2 | Contribution R^2 |
| Percent living in Poverty | Entered | 0.533 | -0.423 |
| Receiving SNAP assistance | Entered |  | -0.355 |
| Black Transformed log10 | Removed |  | -0.12 |
| Hispanic Transformed log10 | Removed |  | 0.124 |
| Median Home value | Removed |  | 0.102 |
| Percent FIRE\_I | Removed |  | 0.050 |
| Blue Collar occupations | Removed |  | -0.071 |
| Workforce Unemployed | Removed |  | -0.099 |
| Two or More Races | Removed |  | -0.225 |
| White | Removed |  | 0.122 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step # 3 | Variable | Entered/removed | Total R^2 | Contribution R^2 |
| Percent living in Poverty | Entered | 0.573 | -0.351 |
| Receiving SNAP assistance | Entered |  | -0.315 |
| Two or More Races | Entered |  | -0.225 |
| Black Transformed log10 | Removed |  | -0.050 |
| Hispanic Transformed log10 | Removed |  | 0.41 |
| Median home value | Removed |  | 0.153 |
| Percent FIRE\_I | Removed |  | 0.081 |
| Blue Collar Occupations | Removed |  | -0.100 |
| Workforce unemployed | Removed |  | -0.011 |
| White | Removed |  | 0.004 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Step # 4 | Variable | Entered/removed | Total R^2 | Contribution R^2 |
| Percent living in poverty | Entered | 0.590 | -0.901 |
| Receiving SNAP assistance | Entered |  | -0.910 |
| Two or More Races | Entered |  | -0.975 |
| Median home value | Entered |  | 0.000 |
| Black Transformed log10 | Removed |  | -0.114 |
| Hispanic Transformed log10 | Removed |  | 0.051 |
| Percent FIRE\_I | Removed |  | -0.008 |
| Blue Collar Occupations | Removed |  | -0.038 |
| Workforce unemployed | Removed |  | -0.053 |
| White | Removed |  | 0.057 |

New:

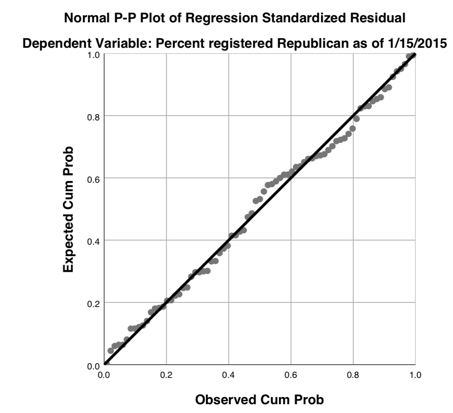
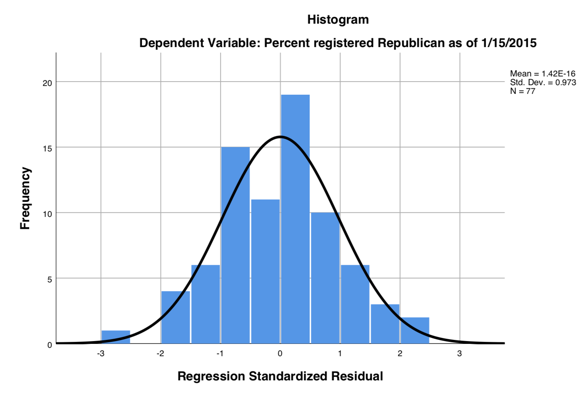
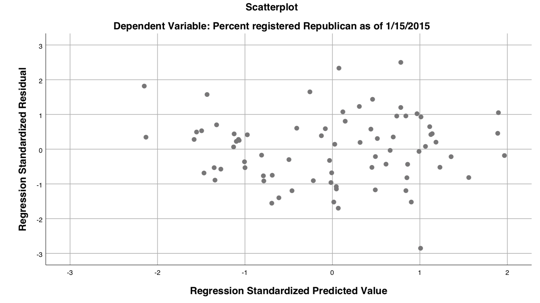
Standard ERROR: 10.02960

F: 25.951

Old:

Standard Error: 10.21705

F: 10.341

* 1. The two variables that I was correct in predicting were going to make it into the model were percent living in poverty, and median home value. I was not correct in guessing that percent of population that identifies as white was going to make it into the final model. The opposite of what I expected to happened with race occurred. Percent two or more races made it into the model instead of white. My thoughts were that percent white help a higher share of the population than percent two or more races and therefore was likely going to represented more than any other races. This was not the case, because percent white never was in any of the models at all steps. Another thing I was not expecting to happen was receiving SNAP assistance making into the final model. There appears to be more influence in receiving SNAP assistance than just income level because that is also described by percent living in poverty.
  2. The first model has the positive side of having more variables than the second model. The more variables drive up the R^2 value, but it can negatively effect the standard error and the significance. In the first model there is an R^2 value of 0.610 and a standard error of 10.22. The second model has a R^2 value of 0.590 and a standard error of 10.03. While the R^2 value is higher in the first model the F statistic shows that the first model is less than half as effective as the second model, 25.951, and 10.341 respectively. The standard error was also improved by 2 tenths from 10.341 to 10.029. This is a key example of parsimony using less to explain the right amount. There are also issues of multicolinearity in the first model that are not as apparent in the second.
  3. 

From the diagnostic graphs generated from SPSS, this model does appear to meet the assumptions of regression in the residuals. This is evidenced in the normal distribution as depicted by the histogram, the homoscedastic pattern in the scatter plot and normality seen in the qq-plot. There might not be complete normality in the model but the general fit to the line on the qq-plot makes it seem like it does.

* 1. The variables I used for this next regression were Percent poverty, SNAP, Median home value, two plus races, Rural, and WestOK.   
       
     The performance of the model is as follows, the model is significant with an R^2 value of 0.606 and an F statistic of 17.95. The standard error is 9.975. The following variables were significant: SNAP (0.013), Poverty (0.013), and Two or more races (0.037). The others that were not were Median home value (0.647), Rural (0.135) and WestOK (0.457). The variables that were negative were: poverty (-0.901), two or more races(-0.851), rural (-5.334), and SNAP (-0.929). The variables that were positive: Median home value (3.555E-5) and WestOK (2.534).
  2. Population by square mile could be another way to measure spatial relations. Possibly changing aggregation to a intermediate scale between school districts and tracts, such as zip codes would also be another way to determine patterns. I believe we could use density as a variable in this study but region may not make much sense nor offer much explanation. In Dense areas ideals can be somewhat more liberal than in less dense areas. Using population density could revel this trend in the data. Region to the extent of my knowledge does not have any direct effect on ones political beliefs. Presumably if there were limited resources within that region and the involved parties wanted to grow it but that logic does not seem to hold up well in long established county lines here in the United States.
  3. Another variable that could be done as a binary is setting a threshold for per captaia income and determine locations as wealth or poor. We could also do a dummy variable based on a threshold of majorly race or not. These two variables would possible offer more differentiation for what is occurring at a case by case basis. For income this may indicate at a glance what is occurring rather than scrutinizing several rows of data to determine if the area is affluent. For the race variable we could determine at a glance which is the majorly.