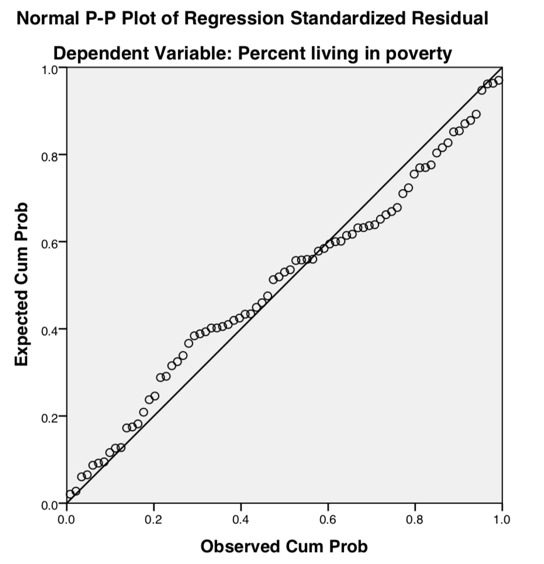
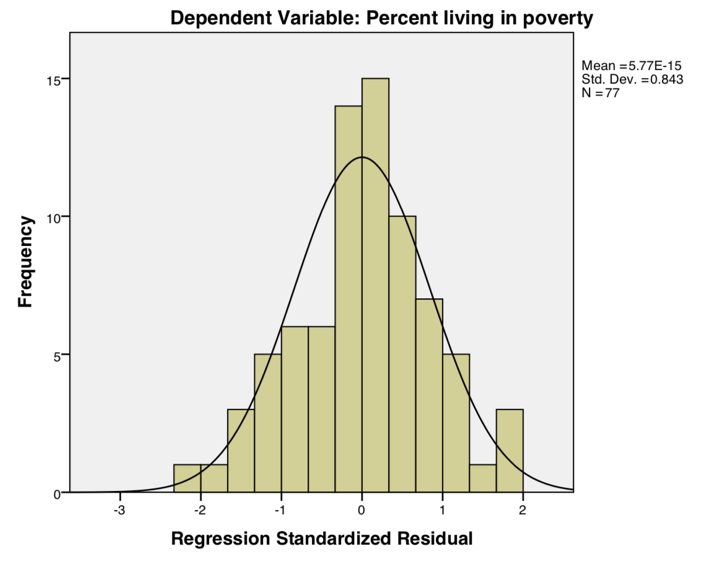
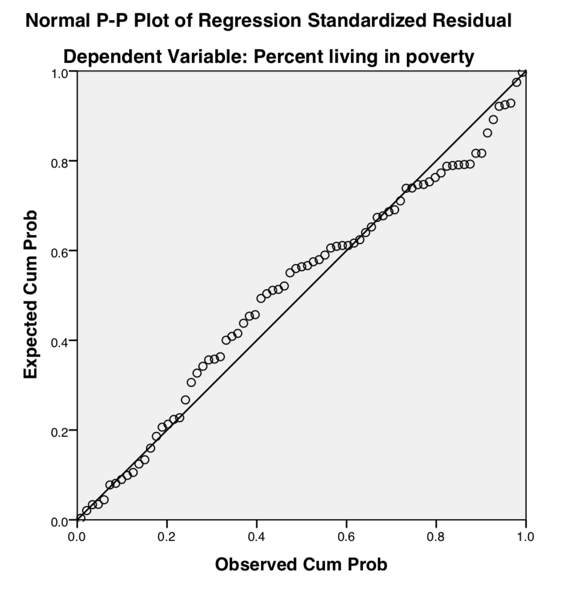
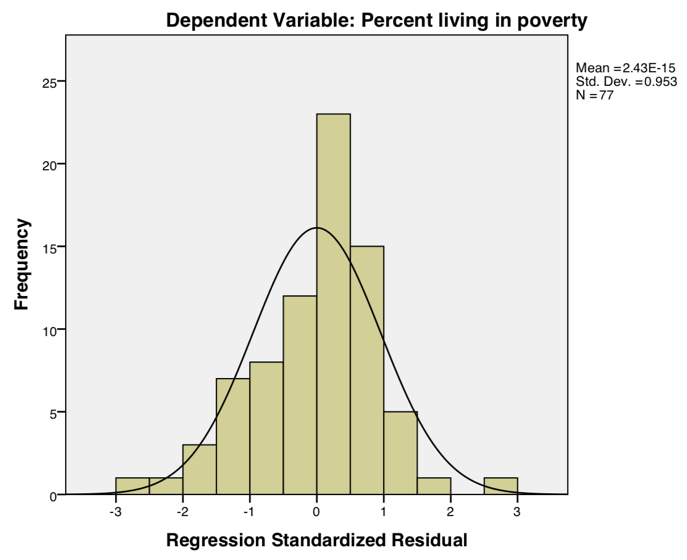
Kellen Bullock

Dr. Comer

3-17-2020

Exercise 5

Part A

* 1.  The kitchen sink model enter regression had a r-square of .829, .759 adjusted. The standard error for this model was 2.2947% poverty by county. This model was also significant with an F score of 11.897. ­Variables that were over the rule of thumb 5 variance inflation factor were: percent other race, percent Hispanic, percent population not in workforce, percent receiving SNAP assistance, per captia income, and percent not completing high school.   
       
       
     This model performs decently well based on the r-square value but there is a good deal of multicolinearlity between the variables as referenced above. In the histogram and the qq-plot we can see there is some deviation from a normal distribution. This does violate another assumption of the regression. This is also reflected in the F score at a pretty low 11.897. Overall this is not a very good model because of the artificial inflation of the r-squared value by the number of variables.   
       
     The variables I predict will be put into the next model are some of the most significant ones in this pervious enter model. They are percent SNAP, and percent vacant housing units. The others I could not determine because of the multicolienarity and the boarder line significance they had in this model.
  2. In the stepwise forward regression model there this a r-square value of .802, .782 adjusted. The standard error of this model is 2.1827% poverty per county. The model is significant with an F score of 39.995. Multicolinearity was not present in this model based on the variance inflation factor. None of the variables were above the 5 VIF threshold. However, there was per capital income that had a VIF of 3.264.   
       
       
       
       
       
       
       
       
       
       
       
       
       
       
       
       
     Above are the plots of the residuals. There are some outliers in this model when observing the histogram. These residuals are not normal and that violates an assumption of the regression. Overall this model is better than the Enter method because the multicolinearity is significantly less at not very much of a cost of the r-square value. However, the residuals are a little worse than in the pervious model in terms of normality. There is however the matter of the F score in this model, which is substantially higher than the pervious one at 39.995.   
       
     The variables that I predicted would make it into this model did from the enter regression kitchen sink. I did not think per captia income would have made it in due to its high multicolienarity with the other variables but it’s VIF is lower in this stepwise model. Percent SNAP and percent vacant housing units did not surprise me when seeing them in this model because in the enter kitchen sink they were already very significant.
  3. The backward stepwise regression had a r-square value of .797, .777 adjusted. The standard error for this model was 2.2093% poverty per county. This was a significant model with an F score of 38.800. The multioclinearity was greatly reduced when compared with the enter model but there were higher VIF variables than in the forwards stepwise regression. These variable did not exceed the rule of thumb 5 VIF and they were Median age, Per captia income, and median home value.   
       
     This model did not perform as well as the forward stepwise regression because the r-square value was lower along with the standard error being greater, and the F score was lower as well. There was also a good bit more multiconlinearity in the backward regression as opposed to the forward regression. I think the backwards one had to iterate more through the variables and possibly missed something that the forward regression found. Since there were 16 iterations in the backward model as opposed to the 7 in the forward.
  4. Per captia income, Percent receiving SNAP assistance, Percent of housing units vacant, Median home value, and Median age are the variables I choose to but into the artisanal model. The artisanal model had an r-square value of .746, .729 adjusted. The standard error of the model was 2.4364% poverty by county. This model was significant and had an F statistic of 41.815. The only variable that was not significant was median home value. Muliconlinearity was not found on the basis of the over 5 rule of thumb but the highest VIF went to per captia income.   
       
     The Artisanal model had a higher F statistic than either the forward or backward stepwise functions. Its strength in explaining was less than either model however with lower r-square value. I do believe this model suffers from less multicolinearity because of the few number of variables, but this is at the cost of explanation that the other two models get form the inclusion of those variables. Unlike the other models there is only 1 variable that isn’t significant which in turn explains why the f statistic is much larger for the artisanal model.
  5. Per Capita Income, Number of Vacant Houses, Percent No highschool diploma, Median House value, and Percent SNAP.   
       
     **Per capita income** seemed like a good variable to choose for SQ777, because groups that would benefit most from this would be wealthy. Advanced technology is expensive as well as the live stock industry, more often than not associated with ranching.   
       
     **Number of Vacant Houses**, would indicate places where both low population density and very stratified income would be where this law could be taken advantage of. Ranching takes a lot of space and in places where land is cheap, usually with abandoned property owners may want this law to pass in order to expand more cost efficiently.   
       
     **Percent No High School Diploma,** seemed like a good choice because those that are less educated, and possibly have less access to high tech, large tracks of land, and are not owners of ranches would be against this law passing.   
       
     **Median House value**, in locations where housing prices are high this law may be advantageous to the home owners. Whereas the lower income, and lower value homeowners would be against it because it would increase the wage gap between the wealthy and the poor.   
       
     **Percent Receiving SNAP Assistance,** this variable was chosen as an indicator of where a combination of low income, low gross domestic product, and low access to resources. The people that fall into this category would likely not want this law to pass because of the further devaluing of what little income they have. It very likely might be they work on these farms and the introduction of new technology would put them out of a job.
  6. The logistic regression for SQ777 had the following significant variables Percent o housing units vacant .001, Median home value .000, and percent not completing high school .006. The other variables that were included in this model were not significant, and they included Per captia income at .453, and Percent receiving SNAP assistance .468.   
       
     The signs of the coefficients were positive for per captia income, which is to be expected the rich were more likely to vote for this bill because as the wealth increases the more likely they can use this law. Percent of housing units vacant was positive as well because in more remote areas that are more suitable to farming. Median home value was also positive and that means the higher the home value the more likely the owners have access to resources to take advantage of this law. Percent receiving SNAP assistance was negative, and this corresponded to the people that have less access to resources are more likely going to vote against this law which enables the ones with resources to use them. Percent not completing high school was positive. This does not make that much sense because usually there are vast income gabs between people with education and people that do not have it.   
       
     Based on the r-square of this model at .510 this model is not good at predicting the outcome of what county will vote in favor or against SQ777. What is dragging the model down is the two variables per captia income and SNAP assistance.
  7. Per Captia Income, Percent No high school diploma, Percent Fire I, percent commute, and Percent republican are the variables I chose.   
       
     **Per captia income,** seemed like a good variable for SQ780 because people that have more access to excess funds would be inclined to shift illegal drug usage to a misdemeanor. If not for themselves but for their children.   
       
     **Percent No High school Diploma,** is a good variable to explain in favor of SQ780, because drug users are more likely to be less educated (dropouts). It is possible this group of people would be in favor of the law because it would help reduce their conviction to prisons.   
       
     **Percent Fire I,** is a group of professionals that includes emergency services. These people interact with a large number of drug users and having their sentences being reduced would likely be the opposite as to what they wanted. Since they would more than likely want the punishment to be as harsh as possible in order to prevent drug users from abusing the law.   
       
     **Percent commute,** this group of people have easy access to transportation and would likely be in favor of this law passing because of their wanting to keep their families from getting into less trouble.   
       
     **Percent Registered Republican,** this group of people would likely be against this law passing because this effect the punishment on commerce. Traditionally both parties have been inclined to stopping or recuing crime rates. Republicans might be more inclined to be one way or the other on this law because it does not discourage stealing, and drug usage.
  8. The logistic regression for SQ780 had the following significant variables, Percent not completing high school, and percent employees in professional industries, The variables that were not significant were Per captia income, percent registered republican as of 1/15/2015, and percent commuting to work alone.   
       
     The signs of the coefficients were positive for per captia income, which makes sense as the wealth of a person increases so too does the wanting to reduces the consequences for their actions. Percent not completing high school did not make sense in the negative direction. This group interesting votes against the reduction of felony charges for the following content of crimes listed in SQ780. I have very much presumed that this group of people would be the ones most actively engaging in these activates. Percent registered republican as of 1/15/2015 was positive much how I predicted. This makes sense because the general stance that republicans have is to reduce the numbers of people in prison. Percent commuting to work alone was negative. This went against my predictions because these people are more likely to have families. It would appear they are against the passing of this law for reasons that escape me.   
       
     Overall this model was pretty bad at determining which counties were likely to vote in favor or against SQ780 because of the low r-square value of .371. The culprits responsible for this low r-square value were the variables that were not significant and they were per captia income, percent not completing high school, percent registered republican, percent commuting to work and percent commuting to work alone.

Part B

* 1. The enter regression for how finished square feet, number of bathrooms, number of fireplaces, lot size, AC, attic, garage, and Age\_Bed does not perform well. The basis for this conclusion is a standard error of $57112.51 per sale price and a low r-square of .551. However, the F statistic is very high at 220.683 and the model is significant.
  2. There is is a little multicolinearity present in this model with the variables Finished square feet (3.232) and Age\_Bed (2.141). These values are not above Rogers’s Rule of thumb 5 VIF. However, there is some similar explanation in the number of bedrooms and the finished square feet of a house.
  3. **Finsihed Square Feet**, has a coefficient in a positive direction at 70.752. This is positive and says the more finished square feet in a house the higher the sold price is. This variable is significant.   
       
     **Number of bedrooms,** has a coefficient of 15101.42 in a positive direction. This means as the number of bedrooms increases the sold price of the home is much large at almost $15,000 per bedroom. This variable is significant at .000.   
       
     **Number of fireplaces,** has a coefficient of 47848.83 in a positive direction. This means there is a increase of home price as the number of fireplaces increases. This variable is significant at .000.  
       
     **Size of Lot**, has a coefficient of .992 in a positive direction. This means as the size of the lot increases so too does that sold home price. This variable is not significant at 0.107.  
       
     **Attic?**, this variable has a coefficient of 18300.30 in a positive direction. This means is if there is an attic the price of a sold house will be higher. This variable is significant at .000.  
       
     **AC?** Has a coefficient of 26817.31 in a positive direction. What this means is if there is a AC the price of the sold house will be higher. This variable is significant at 0.000.  
       
     **Garage?** Has a coefficient of 5389.16 in a positive direction. This means if there is a garage the sale price of the house will be higher. This variable is not significant at 0.315  
       
     **Age\_Bed**, has a coefficient of -74.939, in a negative direction. As the number of bedrooms and age of the house increase the price of the house goes down. This variable is significant at 0.000.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Alderman District | Adj. R2 | St. Error | F test  (p-val) | Sq. Ft  (t-stat) | Lot sz.  (t-stat) | Bed  (t-stat) | Bath  (t-stat) | Fire  (t-stat) |
| 1 | .229 | 22742.979 | 3.074 | 1.518 | 1.441 | -.960 | -.123 | 1.887 |
| 2 | .586 | 20630.892 | 19.670 | 2.215 | 5.194 | -1.417 | 2.174 | -.209 |
| 3 | .830 | 76786.881 | 106.607 | 6.068 | 4.964 | -1.789 | 1.545 | 2.721 |
| 4 | .288 | 66052.872 | 2.012 | 2.300 | .449 | -2.256 | -1.148 | -1.441 |
| 5 | .574 | 27324.099 | 63.862 | 9.389 | 2.171 | -3.702 | 2.124 | 2.265 |
| 6 | .135 | 72998.994 | 1.996 | 1.623 | -.668 | -.838 | -.655 | 1.871 |
| 7 | .226 | 27986.502 | 4.037 | .954 | .535 | -.486 | 2.295 | 1.391 |
| 8 | .216 | 22736.320 | 3.314 | 1.423 | 2.589 | -2.117 | .125 | 1.090 |
| 9 | .426 | 17775.512 | 5.595 | 3.553 | -.209 | -3.809 | .518 | -.994 |
| 10 | .318 | 39425.102 | 18.642 | 1.991 | 4.019 | -.032 | 2.828 | 4.892 |
| 11 | .518 | 26608.486 | 46.507 | 5.834 | .834 | -1.292 | 4.234 | 2.999 |
| 12 | .150 | 32893.566 | 2.016 | 2.044 | -.939 | -1.727 | 1.460 | 0.000 |
| 13 | .571 | 26757.272 | 49.535 | 3.455 | 6.848 | .247 | 2.488 | 3.965 |
| 14 | .385 | 50586.698 | 24.740 | 2.596 | 1.310 | -1.541 | 2.468 | 5.888 |
| 15 | .478 | 32819.917 | 6.30 | -1.998 | .900 | .266 | 3.110 | 3.371 |

The trends that I see between the neighborhoods are fireplaces and finished square feet are the most consistently significant variables. The only neighborhood that is measured adequately from this regression is Alderman District 3 at .830 adjusted r-square. As important as bedrooms are in the overall price regression there is not much significance for them between each neighborhood. It would seem that this quirk behaves differently at this scale. What is very interesting is the highest r-square district had one of the heist standard errors of 76786.881.

* 1. There is definitely spatial autocorrelation occurring in certain districts, where bedrooms, fireplaces, and finished square feet are more important than other areas. The stepwise regression from exercise 4 showed that bedrooms were negative while in the spatial division and here that negative trend is the same in many districts. The overall r-square value for the year was higher for 2 reasons, the higher number of observations, and the larger amount of variables.