GEOGRAPHY 5303
Exercise 5 – 100 points
Due: March 28, 2020, by 5 PM
Extending Regression Analysis

Purpose of Exercise

This exercise continues to analyze socioeconomic indicators in Oklahoma in Part A. We will give Pct_Repub a break as the dependent variable here, and will instead examine poverty as well as the outcomes of two contentious state questions from the November 2016 state election.

Part B continues the exploration of the influences on home sales price in Milwaukee, as with Part A expanding the analysis further forms of regression but looking at spatial rather than temporal subsets of the data.

A. Cross-sectional Social and Demographic Analysis of Oklahoma

A pressing concern in Oklahoma is the poverty level, and this topic receives much attention. You will examine Pct_Poverty as the dependent variable, using regular multivariate OLS regression at the county level only, so select "Scale=C" before commencing your analysis.

1. For the regressions run in a., b., and d. below, you should evaluate the overall model performance, discuss individual variable significance levels including multicollinearity via VIF, and examine residuals (histogram, P-P, and residual plots):
   a. Run a kitchen-sink (selection method: Enter) with all SES variables from Density (Excel column I) through Pct_Repub (column AF) except Pct_White (that is, exclude Pct_White) and assess as indicated above. Also, predict which variables you believe will enter the next regression below.
   b. Run a forward stepwise regression (method: Stepwise) with all the variables indicated in 1.a. and assess the model as assigned, in particular whether you have improved or worsened the residuals. Also, evaluate your prediction accuracy and discuss why or why not you were correct about certain variables.
   c. Run a backward stepwise regression (method: Backward) with all the same variables. Do not assess the whole model, just comment on differences between the forward and backward methods and discuss why you think this happened.
   d. Construct an *artisanal* regression model using a subset of the variables that were in the final models in b. and c. above by considering logic, variable overlap (i.e. VIF), and in general pursuing parsimony by whittling this down to perhaps 4-5 variables you think are most important in predicting poverty. Assess the overall model as before, compare the loss in explanatory power to the models in b. and c., etc.

In addition to other matters that were a foregone conclusion, many state questions were on the November 8, 2016 ballot that received much attention and produced some surprises. Here you will examine two of them, which are produced below in their entirety for your reference:

SQ 777 ("Right to Farm")

The new Section creates state constitutional rights. It creates the following guaranteed rights to engage in farming and ranching:

- The right to make use of agricultural technology
- The right to make use of livestock procedures, and
- The right to make use of ranching practices.

These constitutional rights receive extra protection under this measure that not all constitutional rights receive. This extra protection is a limit on lawmakers' ability to interfere with the exercise of these rights. Under this extra protection, no law can interfere with these rights, unless the law is justified by a compelling state interest-a clearly identified state interest of the highest order. Additionally, the law must be necessary to serve that compelling state interest.

The measure—and the protections identified above—do not apply to and do not impact state laws related to:

- Trespass,
- Eminent domain,
- Dominance of mineral interests,
- Easements
- Right of way or other property rights, and
- Any state statutes and political subdivision ordinances enacted before December 31, 2014

Outcome: 39.7% of total votes in favor, measure failed.

SQ 780 (Reducing punishment for drug crimes)

This measure amends existing Oklahoma laws and would change the classification of certain drug possession and property crimes from felony to misdemeanor. It would make possession of a limited quantity of drugs a misdemeanor. The amendment also changes the classification of certain drug possession crimes which are currently considered felonies and cases where the defendant has a prior drug possession conviction. The proposed amendment would reclassify these drug possession cases as misdemeanors. The amendment would increase the threshold dollar amount used for determining whether certain property crimes are considered a felony or misdemeanor. Currently, the threshold is $500. The amendment would increase the amount to $1000. Property crimes covered by this change include; false declaration of a pawn ticket, embezzlement, larceny, grand larceny, theft, receiving or concealing stolen property, taking domesticated fish or game, fraud, forgery, counterfeiting, or issuing bogus checks. This measure would become effective July 1, 2017.

Outcome: 58.23% of total votes in favor, measure passed (for now).

An Excel file coding each county as 1 if each measure polled over 50% of votes and 0 otherwise is available on D2L, which needs to be added to your current Part A dataset. Just open 5303_EXA.sav, go to the Variable View Tab, add two new Nominal variables (**SQ777** and **SQ780**) with 0 decimals at the bottom of the variable list. Then go to the Data View tab, copy the two columns from Excel (without the column headers), and with the cursor in the first row of the SQ777 column (for Adair County), hit CNTL-V to paste directly into the SPSS data file.

2. SQ777 received the majority of votes in 40 out of 77 counties, but failed by a wide margin statewide.

    a. Choose 5 variables that *you* think might be good predictors, listing each one and explaining why you chose it within the context of the pros and cons of SQ777.

    Next, run a binary logistic regression to predict whether counties would pass SQ777:

    - ☛ Go to "Analyze, Regression, Binary Logistic", put SQ777 in the "Dependent:" box, and your five independent variables in the "Block 1 of 1" window;
    - ☛ Do not change any other settings or options.

    b. Assess the performance of this model via the significance levels of the variables in the equation, whether the signs of the coefficients for each variable make sense, and finally the overall accuracy of the model via the two R Square values computed and the Classification Table. Overall, how good was this model at predicting a county having voted yes or no (in aggregate) on SQ777?

3. SQ780 received the majority of votes in just 31 out of 77 counties, but passed by a wide margin statewide.

    a. Choose 5 variables that *you* think might be good predictors, listing each one and explaining why you chose it within the context of the pros and cons of SQ780.

    Run a binary logistic regression to predict whether counties would pass SQ780.

    b. Assess the performance of this model via the significance levels of the variables in the equation, whether the signs of the coefficients for each variable make sense, and finally the overall accuracy of the model via the two R Square values computed and the Classification Table. Overall, how good was this model at predicting a county having voted yes or no (in aggregate) on SQ780.

B. Home Sales in Milwaukee, 2012

We have learned many things about what impacts home sale prices in Milwaukee as a part of our explorations, but some aspects remain to be considered and will be explored below.

First, assess the possibility of variable interactions in a citywide analysis of sale prices. The weird performance of the number of bedrooms on Exercise 4, being significant but typically having a -$20,000 or so coefficient depending on the regression, is hard to understand, and age of the house was usually insignificant. Perhaps certain ages of houses and their number of bedrooms uncover eras where houses were built a certain way, and thus makes them more or less valuable (i.e. "quaint" or "cozy" older homes with fewer bedrooms, versus modern crap).

- ☛ Go to the Variable View and create a new, "Scale" variable called **Age_Bed**;
- ☛ Under "Transform", click "Compute Variable";
- ☛ Type Age_Bed in the "Target Variable" spot, then type or move variables into the "Numeric Expression" window so it says **Age*Bedrms**.

1. Run a multivariate *enter* regression predicting Sale Price using finished square feet, number of bathrooms, number of fireplaces, lot size, three of the dummy variables (attic, AC, and garage; exclude basement), and the new interaction term as independent variables.
   a. How well does this model predict Y? What is your basis for this evaluation?
   b. Assess the Variance Inflation Factor (VIF) statistic, and evaluate the degree of multicollinearity present in your model.
   c. Discuss each variable; what does the unstandardized coefficient say about the impact of this this trait on sale price, is it significant, etc.? In particular, we are interested in the performance of Age_Bed and interpreting it.

Another factor to consider is whether there is a *spatial* component to home sales prices. We touched on this a little bit when we regressed sale price and included the X and Y coordinates as independent variables, and found that X was significant because it indicated nearness to Lake Michigan, presumably a desirable trait. We will move a little closer to spatial regression here by analyzing "neighborhoods" via the proxy of Alderman district.

In the SPSS data file for Milwaukee, go to "Data", "Split File" and then select "Compare groups" and move Alderman District into the window for "Groups Based on:" Now:

2. Run an *enter* multivariate linear regression predicting sale price with: finished square feet, lot size, number of bedrooms, number of bathrooms, and number of fireplaces. Do not worry about residual plots or multicollinearity statistics this time around. Your output is in the same order as any regression, but each table lists the pertinent statistics for all 15 Alderman districts together for comparison. Using this output, construct a table like is shown below to compare all 15 regressions efficiently, and in the last five columns, put t-statistics (to three decimal points, and including signs!) and then highlight the ones that meet 0.05* and 0.01** significance levels).

You should be able to copy and paste columns from the SPSS output into your Word Table instead of typing them by hand.

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								
2								
Etc.								

a. In aggregate, assess the neighborhood models – what is common across the 15 neighborhoods, what unusual quirks do you see? – discuss anything interesting.
b. Compare these neighborhoods' overall performance to that of the regression you ran on Exercise 4B, Question 3 – stepwise for all 1449 homes using all available variables. Understanding we are not using all variables here, just 5, what patterns do you observe in how well the neighborhood models perform?

What you should upload to Canvas:

- Word document answering all numbered questions above (don't forget to label lettered subparts and to answer all subparts, labeled or otherwise) with SPSS and Excel plots and maps inserted directly where discussed (but no SPSS output *tables* pasted raw – create your own tables in Word when necessary).
- Two SPSS output files (one each for Parts A and B) in **PDF** format.