# GEOGRAPHY 5303
# Exercise 4 – 100 points
# Due: March 7, 2020, by 5 PM
# Relationships between Variables: Multivariate Data

## Purpose of Exercise

This exercise continues to analyze socioeconomic indicators in Oklahoma. Part A of the previous exercise focused strictly on bivariate relationships, examining correlations among the variables first (including partial correlations) and then bivariate regressions between candidate independent variables and the dependent variable, Pct_Repub. Several multivariate regression models are constructed here to more fully explore the influences on voter registration patterns.

Part B continues the exploration of the influences on home sales price in Milwaukee, as with Part A expanding the analysis to multivariate regression but also looking at temporal subsets of the data as assigned to each student previously.

## A. Cross-sectional Social and Demographic Analysis of Oklahoma

At this point it is important that you clearly indicate your pool of ten assigned independent variables (*i.e.*, a list or table in Word). Your choices of $X_1$, $X_2$, $X_3$, and $X_4$ from Exercise 3, Question 3.a. are included by default, but for the six others, make sure to consider transformations (if any) you identified as "better" in Exercise 3 (Question A.1.b.) and used in Question 3. Ultimately, make sure that none of your ten variables is a duplicate of another (*i.e.*, an original and a transform).

1. Regression using the Enter method. Perform a simple linear regression predicting Pct_Repub by using your ten SES variables as independent variables:
   - ☛ Go to "Analyze, Regression, Linear", put Pct_Repub in the "Dependent:" box and your ten independent variables in the "Independents:" window;
   - ☛ In "Statistics", check "Model fit" and "Collinearity diagnostics";
   - ☛ In "Plots", check "Histogram" and "Normal probability plot";
   - ☛ Make sure that the Method window says "Enter", then click "OK".

   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. Also, can you identify specific pairs of variables that strongly correlate? Which ones, and what is your evidence?

   c. Based on results on up to this point, predict the variables you think will or won't wind up in a stepwise model. Explain the reasoning behind your choices. Be honest, and don't do the stepwise regression until after answering this question – correct predictions are not the basis of grading, the logic of your reasoning is.

2. Regression using the Stepwise method. Perform a stepwise regression analysis:
    - ☛ Go to “Analyze, Regression, Linear” change “Enter” to “Stepwise” (make sure all items listed in Question 2 under “Statistics” are still selected);
    - ☛ In “Plots”, make sure “Histogram” and “Normal probability plot” are still selected. Also, set up one scatterplot with *ZRESID on the Y axis and *ZPRED on the X axis (needed for 3.d. below);
    - ☛ In “Save”, select “Distances, Leverage values” (needed for 3.e. below) and “Residuals, standardized” (needed for 3.f. below);
    - ☛ In “Options”, set probability of “Entry” to 0.10 and “Removal” to 0.15.

    a. For each step compile the following: Step #, variable entered/removed, total $R^2$ at each step, and contribution of that variable to total $R^2$ in that step.

    b. Do these results match your predictions from Question 1.c. above and your evaluation from Exercise 3, Question A5.d? If not, what do you think happened that you didn't expect, or failed to account for?

    c. Compare the two regressions you have now performed (Enter and Stepwise) in terms of both the positives and possible negatives encountered.

    d. Assess how well this model meets the regression assumptions vis-a-vis the residuals. Be sure to reference/show all the various statistics/plots generated and what specific assumptions are or are not being met.

    e. Assess the influence of outliers by using Rogerson’s rule of thumb to evaluate the leverage values you have saved. Specify the leverage value that is appropriate here, and then if any counties are influential, identify them and discuss whether their outlier-li-ness is logical.

    f. Insert into Word a classified map of the residuals using a scheme of your choice (as always, explain methods/decisions) and describe any patterns. What are you checking for? What are the implications of not meeting this assumption?

3. Regression using Dummy Variables. You will use population density as a proxy for the rural/urban divide in Oklahoma and also the regions recoded into an east/west split so we can include these as binary (dummy) variables in the analysis:
    - ☛ Go to “Transform”, “Recode into Different Variables”, and move “Population density” in the “Input Variable -> Output Variable:” box;
    - ☛ Click “Old and New Values...”,
    - ☛ Click “Range, LOWEST through value”, enter **54.62** (this is .01 below the state average density), type **1** in “New Value”, “Value” box; and “Add”;
    - ☛ Click “Range, value through HIGHEST”, enter **54.63**, type **0** in “New Value”, “Value” box, click “Add”, and “Continue”;
    - ☛ Type **Rural** in “Output Variable”, “Name” box, click “Change”, “OK”;
    - ☛ Go back to the data file to confirm that a new variable Rural has been created with the expected outcomes, then go to the Variable View tab and set Rural as a “Scale” variable with no decimal points.
    - ☛ Create another new dummy variable called “WestOK” in which regions 3, 4 and 5 are coded with **1** and regions 1, 2, and 6 are coded **0**.

    a. Run a new Enter regression using just the top 3-4 independent variables from Questions 2/3 (state what variables are used) as well as your two new dummy variables. Summarize the results, mostly focusing on the significance of the dummy variables and interpretation of the direction/magnitude of their influence.

b. What other ways could we incorporate the information about regionality and rurality into our regression analysis? Could we simply use Density and Region as independent variables? Why or why not?

c. What other binary traits in Oklahoma might warrant consideration as dummy variables? Why?

4. Run a "Kitchen Sink" multivariate *stepwise* regression predicting Pct_Repub using all 22 *original* SES variables (columns J-AE) and Pct_Fallin (column AG) from the original Exercise 1 Excel file, **plus** the two dummy variables from Question 4, for 25 total independent variables. Use 0.10/0.15 entry/exit levels as assigned on Question 3.

   a. For each step compile the following: Step #, variable entered/removed, total $R^2$ at each step, and contribution of that variable to total $R^2$ in that step.

   b. Compare this regression to the regressions you previously performed on Questions 1 (Enter) and 2 (Stepwise) using just 10 potential independent variables. How much stronger or weaker is this model, what are the significant variables, and overall were there any "important" variables that you were not originally assigned (of your 10)? Did any variables that were significant in your results (Questions 2 and 3) fail to be significant here?

   c. If any dummy variables entered the model, then what is the absolute impact of "Rural" or "WestOK" on voter registrations (from final model, not each step)? Does this make sense?

   d. Holistically, evaluate all regression models run in Questions 1-4 and choose the "best" one, defending your choice with the appropriate information derived from the various analyses, diagnostics, and overall assessment of what we are trying to evaluate in this analysis.

## B. Home Sales in Milwaukee, 2012

On Exercises 1-3 you explored various aspects of the home sale price data, including simple correlations for the entire dataset and subset correlations for each of your three assigned months. Here, you will conduct several regression analyses on the data.

1. Review Exercise 3, Question B.1 and discuss which variables you think are *most* likely to be significant predictors of home sale price, based on the analysis done there.

2. Run a "Kitchen Sink" multivariate *enter* regression predicting Sale Price using all variables pertaining to actual traits of the house: thus, exclude sale date, alderman district, and the X/Y coordinates. Use all 12 months' worth of sales prices.

   ☛ In "Statistics", check "Model fit" and "Collinearity diagnostics";

   ☛ In "Plots", check "Histogram" and "Normal probability plot" and the standard residual scatterplot;

   ☛ In "Save", check "Residuals, standardized";

   ☛ Make sure that the Method window says "Enter", then click "OK".

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. Also, can you identify specific pairs of variables that strongly correlate? Which ones, and what is your evidence?

c. Discuss each significant variable (p-value < 0.05); what does the unstandardized coefficient say about the impact of this this trait on sale price, is the sign logical, etc.? For any particularly unexpected outcomes (at least one exists!), create scatterplot of the suspicious variable (X) versus sale price (Y) and discuss whether/how this plot helps you understand the unexpected outcome.

d. Map the standardized residuals via scatterplot in SPSS. If you want to join your residuals to a shapefile for Milwaukee, you are encouraged to simply classify the residuals as outlined below and map them in GIS. Otherwise, in SPSS recode (reference instructions in Question A4 above) the residuals into five categories:

Category 5 – residuals < -2.0
Category 4 – residuals between -2.0 and -1.0
Category 3 – residuals between -1.0 and 1.0
Category 2 – residuals between 1.0 and 2.0
Category 1 – residuals > 2.0

☛ Create a scatterplot of X and Y, but "Set markers by" the new Residual class variable you created above.

☛ Once the scatterplot has been created, PLEASE rescale the X axis to cover 100,000 units (say, 2,500,000 to 2,600,000) to match the default 100,000 unit span of Y, and also PLEASE PLEASE set decimals for X and Y to 0 (not the defaults of 8/9).

☛ Feel free to play around with different color settings for the five categories if you wish, rather than the defaults selected by SPSS.

Discuss any patterns you detect in the residuals, in particular if there seem to be clusters of (high) positive or (low) negative residuals in certain parts of town.

3. Run a "Kitchen Sink" multivariate *stepwise* regression for all the same variables as in Question B2. Are there any surprises here (the final model) compared to the enter model? If yes, what are they; if not, why not? (You never get away with not having to answer my questions ☺). Write out the full regression equation for the final *stepwise* model, in a format similar to the textbook (*e.g.*, Equations 9.3-9.7 in the text, but use actual (short) variable names instead of $X_1$, $X_2$ – see Equation 9.11).

4. Run a multivariate *stepwise* regression for each of your three assigned months for all house trait variables and compare each month to the model constructed in 3. above:

a. In Word, construct a table for each of the four models (whole year and each of your three months) with adjusted $R^2$, standard error, F/p-value and discuss the variation in model performance between the four models.

b. Are there any notable differences between the year-long model and each of your three months in terms of significant variables? Of great interest would be variables included in the year-long model but not one (or more) of your months, and vice-versa.

## ***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.
- Excel file if needed, otherwise SPSS graphics pasted directly into Word.